FGPO: FEASIBILITY-GATED PREFERENCE OPTIMIZATION FOR VIDEO SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative video super-resolution has achieved remarkable progress in generating perceptually realistic videos. However, existing methods still face key challenges: they struggle to reconcile global structural fidelity with ambiguous local detail recovery, often yielding over-smoothed textures or geometric distortions. Moreover, preference- and reinforcement learning-based approaches lack spatially differentiated supervision, increasing the risk of structural artifacts while offering limited guidance for detail synthesis. To address these issues, we propose **FGPO**, a feasibility-gated preference optimization framework for video super-resolution. FGPO first integrates multi-objective evaluation and feasibility-gated Pareto selection, utilizing LR-conditioned detail plausibility and temporal consistency scores to restrict candidate eligibility and prevent reinforcing low-fidelity outputs. It then performs LR-guided spatiotemporal dispersion estimation and mask generation, aligning candidate dispersion with observable LR structures to isolate regions with high uncertainty. Finally, a dispersion-weighted inter-candidate feature diversity bonus is introduced to encourage localized detail variation in high-dispersion areas while preserving low-dispersion LR-aligned structures. FGPO provides update-rule-compatible scalar and pairwise feedback, enabling seamless compatibility with various post-training algorithms. Extensive experiments demonstrate that FGPO achieves state-of-the-art performance across reconstruction fidelity, temporal consistency, and perceptual quality, consistently outperforming existing approaches on multiple benchmarks.

1 INTRODUCTION

Recent advances in deep neural architectures and generative priors have substantially improved the perceptual quality and visual realism of video super-resolution (VSR) (Wang et al., 2019; Chan et al., 2022a; Xu et al., 2025; Zhou et al., 2024; Xie et al., 2025). Nevertheless, VSR remains a spatially and temporally non-uniform inverse problem because the LR observation constrains different regions with varying degrees of local identifiability (Chan et al., 2021; 2022a). At the macroscopic scale, a low-resolution (LR) video preserves coarse structural cues and partially informative motion trajectories that constrain the corresponding high-resolution (HR) reconstruction. However, at microscopic scales, downsampling, motion blur, and video compression severely degrade high-frequency evidence in regions such as hair, grass, and facial features. Consequently, regions with high local identifiability admit relatively constrained HR reconstructions, whereas regions with low local identifiability may correspond to multiple plausible high-frequency details. Conventional VSR models typically optimize spatially uniform reconstruction objectives without explicitly accounting for these regional differences (Wang et al., 2019; Chan et al., 2021). As a result, minimizing the expected reconstruction error in regions with low local identifiability encourages predictions toward averaged solutions, leading to over-smoothed textures and weak fine details (Ledig et al., 2017; Wang et al., 2018; Blau & Michaeli, 2018). This limitation reflects a mismatch between spatially varying local identifiability and globally uniform training objectives.

The learned perceptual rewards and preference optimization provide an effective way to improve perceptual quality in generative restoration (Xu et al., 2023; Yi et al., 2026). However, the admissible local variation of VSR reconstructions varies with local identifiability across space and time: regions with high local identifiability should remain faithful to LR-supported structures and temporally consistent, whereas regions with low local identifiability admit limited yet beneficial fea-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

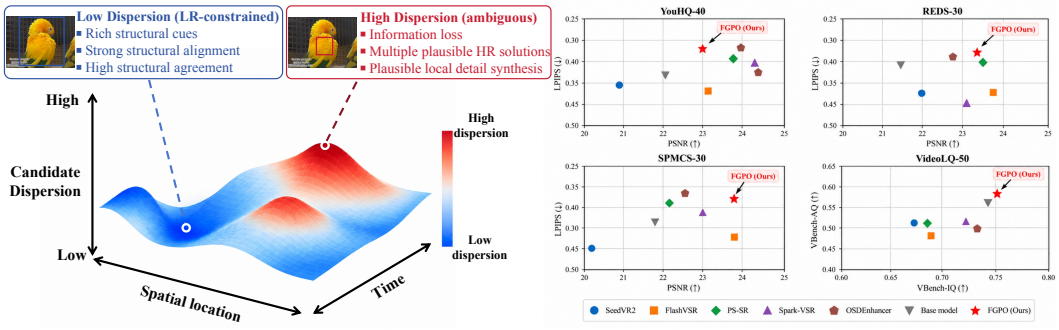

Figure 1: FGPO estimates spatially and temporally varying candidate dispersion in video super-resolution, assigning lower reward weights around LR-aligned structures while encouraging local feature variation in high-dispersion regions.

ture exploration within the constraints of the observed LR input. Existing reinforcement learning methods typically rely on video-level or sample-level rewards without explicit spatiotemporal localization. Consequently, perceptual optimization may improve local textures at the cost of structural fidelity in well-constrained regions, while providing insufficient guidance for recovering plausible details in poorly constrained regions. We refer to this discrepancy as the mismatch between spatially varying local identifiability and globally uniform optimization signals.

To address this limitation, we formulate preference optimization for VSR through a hierarchical feedback construction. First, candidate videos are partitioned into feasible and infeasible subsets using LR-conditioned detail-plausibility and temporal-consistency criteria, reducing the risk of reinforcing structurally inconsistent outputs. Second, within the gate-passing feasible set, first-front Pareto selection over clarity and aesthetic quality identifies perceptually competitive candidates, establishing eligibility for an additional feature-discrepancy bonus. Finally, a dispersion-weighted inter-candidate feature diversity bonus is restricted to Pareto-eligible candidates and spatially modulated by an LR-guided dispersion mask, encouraging local feature variation in high-dispersion regions while assigning lower dispersion weights around LR-aligned structures.

Based on this formulation, we propose **FGPO**, a Feasibility-Gated Preference Optimization framework for VSR that constructs reliable preference signals through feasibility-gated Pareto selection and dispersion-weighted reward modeling. Specifically, FGPO integrates multi-objective evaluation with feasibility-gated Pareto selection, using feasibility scores to restrict candidate eligibility and identify perceptually non-dominated outputs. It then leverages LR-guided spatiotemporal candidate-dispersion estimation to align empirical dispersion with observable LR structures, generating a soft spatial mask. Finally, a dispersion-weighted inter-candidate feature diversity bonus encourages local feature variation in high-dispersion regions while assigning lower reward weights around LR-aligned structures. FGPO exposes scalar and pairwise feedback interfaces, enabling seamless compatibility with representative preference optimization algorithms such as GRPO (Liu et al., 2026a) and DPO (Wallace et al., 2024). Extensive experiments demonstrate that FGPO achieves a competitive trade-off among reconstruction fidelity, temporal consistency, and perceptual quality.

The main contributions of this paper are summarized as follows:

- We formulate VSR preference optimization via a hierarchical feedback framework, introducing a feasibility-gated Pareto-front eligibility mechanism that prioritizes structural plausibility and temporal consistency before determining eligibility for discrepancy rewards.
- We develop an LR-guided candidate-dispersion estimation mechanism paired with a dispersion-weighted inter-candidate feature diversity bonus, enabling localized detail enhancement while suppressing ungrounded variations around LR-supported structures.
- FGPO serves as an update-rule compatible preference-feedback framework compatible with representative group-relative and pairwise preference-optimization algorithms, achieving state-of-the-art performance across multiple benchmarks.

2 Related Work

Video Super-Resolution. Early methods in video super-resolution primarily formulated the problem as deterministic regression, such as DUF (Jo et al., 2018), EDVR (Wang et al., 2019), BasicVSR (Chan et al., 2021), BasicVSR++ (Chan et al., 2022a), VRT (Liang et al., 2024), and RVRT (Liang et al., 2022), have advanced VSR performance through spatial alignment, recurrent propagation, and long-range temporal modeling, respectively. However, regression objectives inherently force these methods to optimize for the conditional mean. As a result, they yield over-smoothed textures in ambiguous regions where low-resolution cues cannot uniquely constrain high-resolution details. Generative VSR methods (Chu et al., 2020; Chan et al., 2022b; Zhou et al., 2024; Xie et al., 2025) address this limitation by integrating adversarial networks (Goodfellow et al., 2020) or diffusion models (Ho et al., 2020). Particularly, approaches leveraging large-scale pretrained generative models—such as SeedVR2 (Wang et al., 2026a), FlashVSR (Zhuang et al., 2026), Vivid-VR (Bai et al., 2025), OSDEnhancer (Wei et al., 2026), PS-SR (Wu et al., 2026), and SparkVSR (Yu et al., 2026)—demonstrate superior capabilities in detail synthesis, degradation adaptation, and reference-free quality guidance. However, existing generative VSR methods allow generation uniformly across the frame, failing to balance LR-guided structural regions with regions where high-frequency information is irrecoverably lost. To address this, FGPO dynamically allocates feature-exploration incentives based on candidate feasibility and local spatiotemporal dispersion, achieving a favorable balance between LR-supported structural fidelity and plausible detail synthesis.

Reinforcement Learning. Reinforcement learning (RL) provides a general framework for optimizing generative models against task-specific or human-preference reward signals, and has been widely adopted in both language models (Schulman et al., 2017; Guo et al., 2025; Rafailov et al., 2023; Ouyang et al., 2022) and vision models (Black et al., 2024; Liu et al., 2026a;b; Wang et al., 2026b; Xu et al., 2023). Particularly in large-scale industrial models (Zhao et al., 2026; Diao et al., 2026; Cai et al., 2025; Seedream et al., 2025), post-training with RL significantly enhances model capabilities. In visual super-resolution, RL-based methods (Yi et al., 2026; Qiao et al., 2025; Liu et al., 2026c) also gain popularity to improve perceptual quality and human-preference alignment. However, tailored for image super-resolution, these methods predominantly rely on spatially indiscriminate rewards. This not only induces geometric distortions in structured regions and temporal flickering, but also fails to encourage local feature variations in blurry areas, impeding RL exploration and ultimately leading to sub-optimal performance in video super-resolution tasks. To address these limitations, our feasibility-gated preference optimization promotes structural and temporal consistency through a gated candidate partition with violation-dependent feedback, and concentrates feature exploration on local spatiotemporal high-dispersion regions via a soft LR-guided weighting mask.

3 Methodology

We propose FGPO, an update-rule-compatible preference optimization framework formulating VSR as a constrained multi-objective problem. In Fig. 2, FGPO partitions candidates by plausibility and temporal consistency, identifies Pareto-optimal candidates within the feasible set, and builds an LR-guided dispersion mask to spatially weight feature-discrepancy rewards. The feedback is provided as sample-level scalar rewards or pairwise preferences, without prescribing downstream update rules.

3.1 Multi-Objective Evaluation and Feasibility-Gated Pareto Selection

Given a low-resolution video y, we inject K noise samples $\{z_i\}_{i=1}^{K}$ into the latent space to generate K high-resolution candidate videos $\{x_i\}_{i=1}^{K}$. To build a task-specific evaluator, we construct a multi-objective instruction-tuning dataset based on RealVSR (Yang et al., 2021). Specifically, Gemini-3.8-Flash (Google DeepMind, 2026) generates 5,000 quality annotations for each of the four evaluation dimensions. To mitigate potential unaligned preferences or biases from the AI evaluator, experienced human annotators double-checked these annotations. We then fine-tune Qwen3.5-Omni (Team, 2026) on these annotations, enabling online feedback during reinforcement learning, as detailed in Appendix B. For each candidate x_i, the fine-tuned evaluator produces four normalized scores $q_i = [q_i^D, q_i^T, q_i^C, q_i^A]$, corresponding to detail plausibility, temporal consistency, spatial clarity, and aesthetic quality. These scores form the evaluation matrix $Q = [q_1; \ldots; q_K] \in \mathbb{R}^{K\times 4}$.

We treat LR-conditioned detail plausibility and temporal consistency as operational feasibility criteria. Within the feasible subset, candidates are evaluated according to spatial clarity and aesthetic

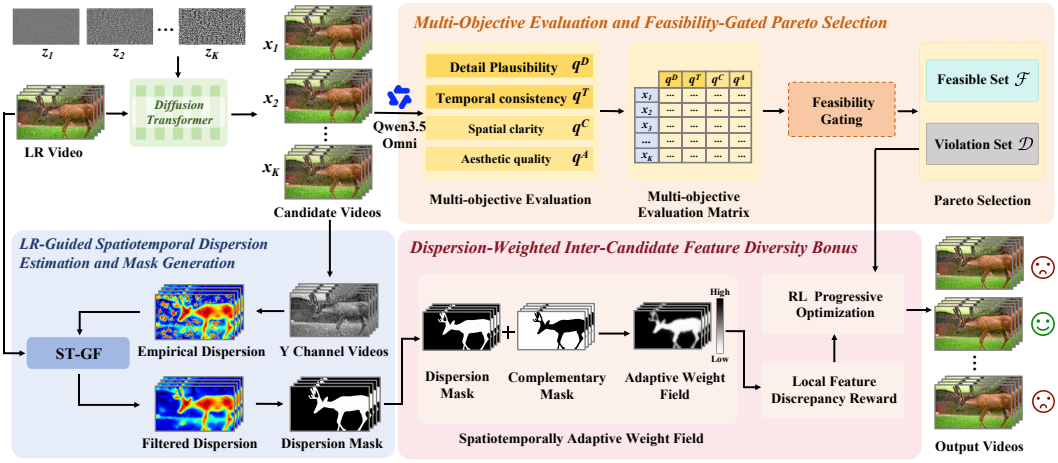

Figure 2: **Overview of FGPO.** FGPO partitions VSR candidates using feasibility criteria, applies Pareto selection to establish eligibility for feature-discrepancy rewards, and uses an LR-aligned candidate-dispersion mask to weight local rewards, delivering scalar or pairwise feedback for downstream post-training algorithms.

quality, while local detail variation is governed in a subsequent optimization stage. To implement this hierarchy, we construct a two-channel feasibility gate using the corresponding dimensions of Q:

$$g_i^D = \sigma\big(\beta_D(q_i^D - \tau_D)\big)\,, \qquad g_i^T = \sigma\big(\beta_T(q_i^T - \tau_T)\big)\,, \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function, g_i^D and g_i^T are the detail plausibility and temporal consistency gate values, τ_D and τ_T are predetermined thresholds uniquely bounding the hard feasible set ($q_i \geq \tau \iff g_i \geq \sigma(0)$), and $\beta_D, \beta_T > 0$ control the steepness of downstream soft quality modulation. Threshold calibrations, pass rates, boundary analyses are in Appendix H and L. Thresholding the gate $g_i = [g_i^D, g_i^T]^\top \in [0,1]^2$ partitions the candidate videos into the gate-passing feasible set $\mathcal{F}$:

$$\mathcal{F} = \big\{x_i \,\big|\, g_i^D \geq \sigma(0) \;\wedge\; g_i^T \geq \sigma(0)\big\}\,, \qquad \mathcal{D} = \{x_i \mid x_i \notin \mathcal{F}\}, \tag{2}$$

where $\mathcal{D}$ contains candidates that fall below at least one feasibility threshold. Candidates in $\mathcal{D}$ are excluded from positive perceptual and local-diversity optimization, receiving violation-dependent negative feedback in the subsequent preference-optimization stage.

For candidates in $\mathcal{F}$, we perform Pareto selection with respect to spatial clarity and aesthetic quality. Let $\mathrm{ND}(\cdot)$ denote the first non-dominated front under the maximization of q^C and q^A, then the feasible Pareto set is defined as $\mathcal{P} = \mathrm{ND}\left(\mathcal{F}; q^C, q^A\right)$. When $\mathcal{F} = \varnothing$, Pareto analysis is skipped ($\mathcal{P} = \varnothing$) and all candidates fall into $\mathcal{D}$; if $|\mathcal{F}| = 1$, it yields $|\mathcal{P}| = 1$ (where $R_i^{\mathrm{L}} = 0$); otherwise, the non-dominated front is non-empty. Feasibility gating establishes candidate-level filtering, complemented by spatial reward modulation in Sec. 3.3 for fine-grained local feature regulation. Pareto front dynamics, discrete score ties, and stability analyses are detailed in Appendix K.

3.2 LR-Guided Spatiotemporal Dispersion Estimation and Mask Generation

Following candidate partitioning, we estimate an empirical candidate-dispersion proxy using all K candidates. Because all generated outputs are strongly constrained by the deterministic LR input y, they preserves consistent global geometry, making full-cohort dispersion an informative spatial signal for local generative ambiguity rather than structural collapse. We convert candidates $\{x_i\}_{i=1}^K$ to the YCbCr color space (Poynton, 2012) and extract luminance channels $x_i^Y \in \mathbb{R}^{T\times H\times W}$, which emphasizes structural variation while reducing sensitivity to chromatic fluctuations.

We compute the empirical sample-dispersion field $V = \frac{1}{K}\sum_{i=1}^K \left(x_i^Y - \bar{x}^Y\right)^2 \in \mathbb{R}^{T\times H\times W}$, where $\bar{x}^Y = \frac{1}{K}\sum_{i=1}^K x_i^Y$ and operations are element-wise. To spatially regularize the dispersion field along LR-observable boundaries, we use the luminance channel $y^{\uparrow,Y}$ of the bicubically upsampled LR video as guidance. Extending guided filtering (He et al., 2012) to spatiotemporal neighborhoods

(ST-GF), we obtain $\tilde{V} = \text{ST-GF}\left(V, y^{\uparrow,Y}; r, \epsilon\right)$, where r specifies spatiotemporal neighborhood radii, and ϵ regularizes local linear fitting. ST-GF smooths local fluctuations while suppressing leakage across deterministic LR boundaries, anchoring spatial weighting strictly to LR-aligned uncertain regions (see Appendix N and O for extended mask validations and comparisons).

Because finite candidates provide only a sample-based estimate, we interpret $\tilde{V}$ as relative empirical dispersion rather than a parametric posterior variance. To reduce video-dependent scale variation, we normalize it by its global spatiotemporal mean $\mu_{\tilde{V}} = \sum \tilde{V}/(T \cdot H \cdot W)$ with stabilization constant $\delta > 0$ as $d = \tilde{V}/(\mu_{\tilde{V}} + \delta)$. Finally, we map relative dispersion to a soft spatiotemporal mask $M = 1/(1 + \exp[-\alpha(d-1)])$, where $\alpha > 0$ controls transition sharpness around the video-wide average. The resulting mask $M \in [0,1]^{T \times H \times W}$ assigns larger weights to regions with above-average LR-aligned empirical candidate dispersion and smaller weights to regions with lower relative dispersion.

3.3 Dispersion-Weighted Inter-Candidate Feature Diversity Bonus

While the candidate-level feasibility gate filters out grossly inconsistent outputs, localized artifacts may still persist within gate-passing candidates. To mitigate localized deviations during optimization, we introduce a dispersion-weighted inter-candidate local feature diversity bonus for candidates in $\mathcal{P}$. This reward encourages feature-space differences across stochastic samples, while guiding feature-discrepancy incentives primarily toward regions with higher empirical candidate dispersion.

We construct a spatiotemporally adaptive weight field $W_{\text{F}} = w_{\text{det}}(1 - M) + w_{\text{stoch}} M$ from the LR-guided mask, where $0 \leq w_{\text{det}} < w_{\text{stoch}}$. Owing to the normalization denominator in Eq. 3, the relative ratio $w_{\text{stoch}}/w_{\text{det}}$—rather than their absolute scale—governs the spatial focus of feature feedback. Specifically, low-dispersion regions receive downweighted relative importance via w_{det}, while high-dispersion regions are prioritized via w_{stoch} to encourage feature-space exploration in regions with higher empirical candidate dispersion.

For each $x_i \in \mathcal{P}$, we extract spatiotemporal features f_i using a frozen feature encoder (Tong et al., 2022) and normalize them along channel dimension, obtaining $\hat{f}_i$. For a candidate pair (x_i, x_j), the position-wise cosine-distance field D_{ij} and its dispersion-weighted aggregation $\bar{D}_{ij}$ are defined as:

$$D_{ij} = \frac{1 - \langle \hat{f}_i, \hat{f}_j \rangle_c}{2}, \qquad \bar{D}_{ij} = \frac{\langle D_{ij}, W_{\text{F}} \rangle}{\sum W_{\text{F}} + \epsilon_{\text{F}}}, \tag{3}$$

where $\langle \cdot, \cdot \rangle_c$ denotes the inner product over feature channels, and the second inner product sums element-wise products across all spatiotemporal positions. Since features are ℓ_2-normalized, $D_{ij} \in [0,1]$ and $\bar{D}_{ij} \in [0,1]$. Let $\mathcal{N}_k(i)$ denote the k nearest neighbors of x_i in $\mathcal{P}$ according to $\bar{D}_{ij}$, where $k = \min(k_0, |\mathcal{P}| - 1)$. We define the inter-candidate local feature diversity bonus as $R_i^{\text{L}} = \sum_{j \in \mathcal{N}_k(i)} \bar{D}_{ij}/k$ when $|\mathcal{P}| \geq 2$, and $R_i^{\text{L}} = 0$ when $|\mathcal{P}| = 1$. Using local k-nearest neighbors and bounded cosine distances prevents isolated structural outliers from inflating rewards, ensuring that diversity is encouraged only among plausible, non-dominated candidates. To integrate feasibility and perceptual quality, we define the base quality reward R_i^{Q}:

$$R_i^{\text{Q}} = w_D q_i^D + w_T q_i^T + g_i^D g_i^T \left(w_C q_i^C + w_A q_i^A\right), \tag{4}$$

where $w_D, w_T, w_C, w_A \geq 0$ weigh the four evaluation dimensions. The final positive feedback for a feasible Pareto candidate $x_i \in \mathcal{P}$ is defined as $R_i^{\text{P}} = R_i^{\text{Q}} + R_i^{\text{L}}$.

This two-tier design pairs candidate-level feasibility partitioning with spatial reward modulation, achieving fine-grained feature regulation that discourages low-scoring candidates. Via LR-guided spatial filtering, it directs feature-space discrepancies toward empirically high-dispersion regions, thereby mitigating localized artifacts. The local discrepancy reward is active only during training; at inference, fixing the random seed with a deterministic sampler yields reproducible reconstructions.

3.4 Update-Rule-Compatible Feasibility-Gated Optimization

FGPO is formulated as an update-rule-compatible feedback construction framework, rather than a specific policy-update algorithm. For an infeasible candidate $x_i \in \mathcal{D}$, we define its normalized detail-plausibility and temporal-consistency violations as $v_i^D = [\tau_D - q_i^D]_+/s_D$ and $v_i^T = [\tau_T - q_i^T]_+/s_T$, respectively, where $[u]_+ = \max(u, 0)$, and assign violation-dependent negative feedback:

$$R_i^{\text{V}} = -\lambda_D \rho(v_i^D) - \lambda_T \rho(v_i^T) - \lambda_{DT} \rho(v_i^D v_i^T), \qquad \rho(u) = \log(1 + u). \tag{5}$$

Together with the rewards in Sec. 3.3, this yields the unified feasibility-aware scalar feedback $\widetilde{R}_i$, where $\widetilde{R}_i = R_i^{\mathrm{P}}$ for $x_i \in \mathcal{P}$, R_i^{Q} for $x_i \in \mathcal{F} \setminus \mathcal{P}$, and R_i^{V} for $x_i \in \mathcal{D}$. Pareto membership serves as an eligibility rule for the feature-discrepancy bonus rather than defining a full preference ordering. Candidates in $\mathcal{P}$ receive R_i^{P}, whereas other gate-passing candidates in $\mathcal{F} \setminus \mathcal{P}$ retain R_i^{Q}. Thus, a candidate in $\mathcal{F} \setminus \mathcal{P}$ with a sufficiently high base score R_i^{Q} may receive a larger total reward than one in $\mathcal{P}$. This soft hierarchy restricts diversity rewards while preserving non-front learning signals.

Given candidates sampled for the same LR video, FGPO maps each candidate to the scalar feedback in $\widetilde{R}_i$. This construction depends solely on generated candidates, the LR condition, and evaluators; it is independent of downstream advantage estimators, policy-ratio parameterizations, clipping strategies, or model update rules. The same feedback can also be converted into preference supervision for pairwise optimization. Specifically, candidates are ordered by the feasibility hierarchy, preferring feasible candidates over infeasible ones, and ranked within the feasible set based on quality and dispersion-weighted discrepancy rewards. FGPO exposes two standard feedback interfaces:

$$\mathcal{I}_{\mathrm{scalar}} = \left\{ \widetilde{R}_i \right\}_{i=1}^{K}, \qquad \mathcal{I}_{\mathrm{pair}} = \left\{ \left(x_w^{(m)}, x_l^{(m)} \right) \right\}_{m=1}^{M}, \tag{6}$$

where $\mathcal{I}_{\mathrm{scalar}}$ serves reward-based optimization, $\mathcal{I}_{\mathrm{pair}}$ serves pairwise preference optimization, and M denotes the number of valid preference pairs. The candidate group is required to estimate relative empirical dispersion and construct feedback, without prescribing downstream update rules.

As representative instantiations, we integrate FGPO with GRPO and DPO. For GRPO, scalar feedback is converted into a group-relative advantage $A_i = (\widetilde{R}_i - \mu_{\widetilde{R}})/(\hat{\sigma}_{\mathrm{trim}} + \epsilon_0)$, where $\hat{\sigma}_{\mathrm{trim}}$ is the robust trimmed deviation estimator (Appendix N) used in the clipped policy objective. For DPO, the feasibility hierarchy and $\widetilde{R}_i$ construct winner–loser pairs optimized via the standard reference-relative preference objective. These instantiations modify only the feedback adapter while leaving the feasibility criterion, Pareto selection, candidate-dispersion estimation, and local diversity reward unchanged.

4 Experiments

Datasets and Evaluation Metrics. We evaluate FGPO on diverse benchmarks covering both synthetic and real-world video degradations. The synthetic datasets include REDS-30 (Nah et al., 2019), SPMCS-30 (Tao et al., 2017) with various downscaling factors, and YouHQ-40 (Zhou et al., 2024) containing 1080p high-definition videos. The real-world datasets encompass VideoLQ-50 (Chan et al., 2022b) with complex natural degradations. We evaluate reconstruction fidelity using PSNR and SSIM, and perceptual quality using LPIPS (Zhang et al., 2018). Temporal consistency is measured by VBench-SC (Huang et al., 2024), which assesses subject consistency across frames. We also report VBench-IQ for image quality and VBench-AQ (Huang et al., 2024) for aesthetic quality.

Implementation Details. Following FlashVSR (Zhuang et al., 2026), we establish our Base Model by adapting the open-source video generation model Wan2.1-1.3B (Wan et al., 2025) for video super-resolution. Following OSDEnhancer (Wei et al., 2026), the training data for the Base Model is constructed with HQ-VSR (Chen et al., 2026), Adobe240 (Su et al., 2017), and DIV2K (Cai et al., 2019) datasets. On top of this Base Model, we apply the proposed FGPO preference-alignment protocol, where a 20% subset of the dataset is sampled for the GRPO training phase. In the default FGPO training stage, the candidate group size is set to 16, the learning rate is 1×10^{-6}, and the KL-divergence penalty coefficient is 0.05. The evaluation matrix weights are set to $w_D = 1.0, w_T = 1.0, w_C = 0.5, w_A = 0.5$, and the feasibility gating activation thresholds are set to $\tau_D = 0.15, \tau_T = 0.85$. During inference, the random seed is fixed to 42 for evaluation. Detailed hyperparameter summaries are provided in Appendix E.

4.1 Quantitative Results

Table 1 compares FGPO with representative VSR methods and Base Model across four benchmark datasets. On the synthetic benchmarks YouHQ-40 (Zhou et al., 2024), REDS-30 (Nah et al., 2019), and SPMCS-30 (Tao et al., 2017), FGPO consistently elevates the Base Model and outperforms all existing VSR approaches across all metrics, including reference-based fidelity metrics (PSNR, SSIM, LPIPS) (Zhang et al., 2018) and non-reference perceptual dimensions evaluated by

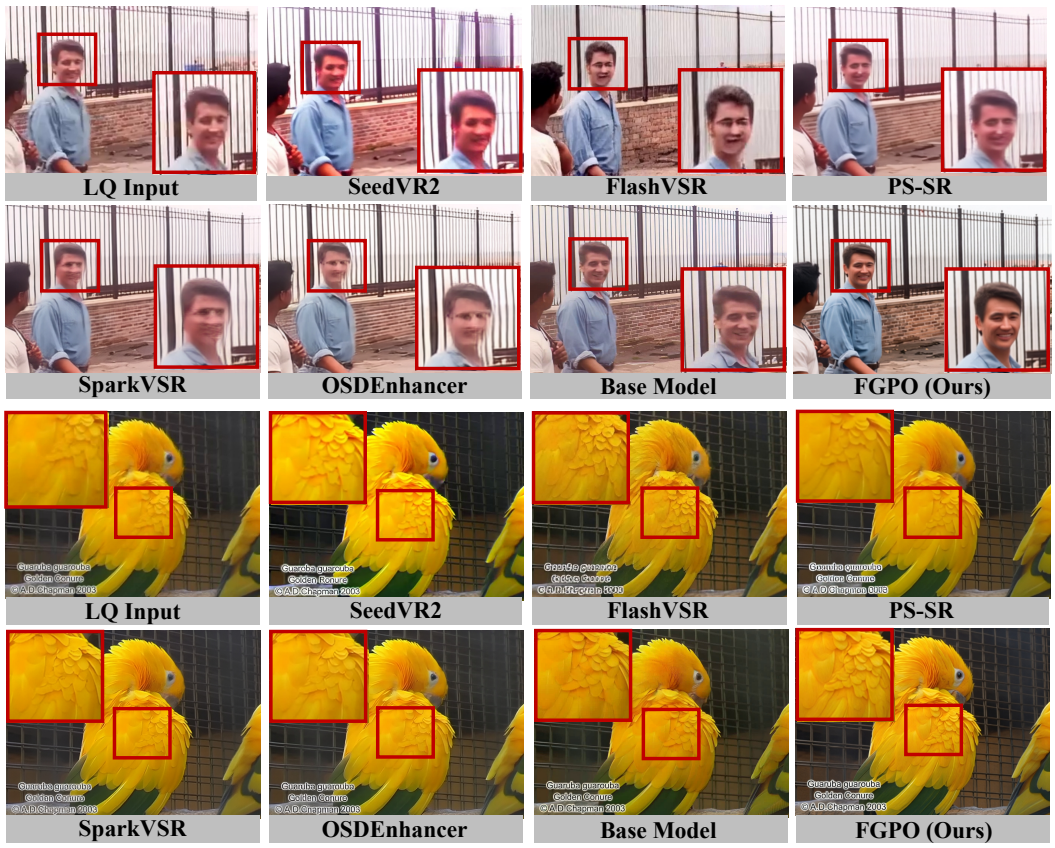

Figure 3: **Qualitative comparison with state-of-the-art VSR methods.** FGPO produces sharper local textures with fewer visible changes to coarse object geometry compared to existing methods.

VBench (Huang et al., 2024). Unlike existing methods that often trade perceptual realism for pixel-level accuracy, FGPO achieves state-of-the-art performance in both reconstruction fidelity and visual quality. Notably, on the unreferenced real-world VideoLQ-50 (Chan et al., 2022b) benchmark, FGPO likewise achieves optimal performance across all VBench metrics, indicating strong robustness under complex real-world degradations. These experimental results benefit from FGPO's hierarchical framework, where candidate-level feasibility partitioning reduces the risk of structural distortions before spatially guided feature discrepancy optimization promotes plausible detail synthesis. As presented in Appendix F, FGPO achieves perceptual quality closest to real videos. Furthermore, comprehensive quantitative evaluations of other framework variants are detailed in Appendix C.

4.2 Comparison with Different Training Strategies

Table 2 compares FGPO with alternative training strategies. Standard supervised fine-tuning enhances pixel-level fidelity but compromises perceptual quality, as optimizing toward a single deterministic target fails to encourage plausible high-frequency details. Directly applying unconstrained post-training methods such as vanilla GRPO or DPO suffers from fidelity degradation and limited perceptual gains; without explicit feasibility bounds, preference optimization incorrectly favors visually sharp candidates that suffer from structural distortions or temporal jitter. Global-feedback variants similarly fail to achieve a balanced trade-off, as applying a spatially uniform reward can introduce artificial variation into regions that exhibit relatively low candidate dispersion under the LR condition. In contrast, FGPO resolves these conflicts through structured feedback construction. Feasibility gating partitions out candidates that fail learned plausibility criteria, Pareto selection determines eligibility for feature-discrepancy rewards among gate-passing candidates, and the LR-guided mask directs feature-discrepancy rewards toward empirically high-dispersion regions. Consequently, FGPO achieves simultaneous gains in all fidelity, consistency, and perceptual metrics over the Base Model, validating the necessity of spatially selective and feasibility-aware feedback.

Table 1: Quantitative comparison on four video super-resolution benchmarks. The best and second-best results are highlighted in red and orange, respectively.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	VBench-SC ↑	VBench-IQ ↑	VBench-AQ ↑
YouHQ-40	SeedVR2	21.167	0.632	0.376	0.948	0.630	0.526
	FlashVSR	23.130	0.604	0.387	0.975	0.745	0.580
	PS-SR	23.893	0.669	0.312	0.978	0.730	0.569
	SparkVSR	24.443	0.686	0.343	0.975	0.725	0.568
	OSDEnhancer	24.322	0.676	0.289	0.974	0.735	0.548
	Base Model	22.255	0.594	0.364	0.970	0.732	0.565
	FGPO (Ours)	24.562	0.702	0.275	0.979	0.762	0.592
REDS-30	SeedVR2	22.192	0.599	0.429	0.839	0.526	0.461
	FlashVSR	23.920	0.649	0.419	0.896	0.741	0.515
	PS-SR	23.552	0.621	0.324	0.901	0.679	0.475
	SparkVSR	23.196	0.594	0.454	0.900	0.616	0.483
	OSDEnhancer	22.936	0.587	0.334	0.896	0.714	0.473
	Base Model	21.549	0.535	0.365	0.890	0.751	0.504
	FGPO (Ours)	24.035	0.662	0.310	0.902	0.768	0.535
SPMCS-30	SeedVR2	19.735	0.539	0.447	0.887	0.502	0.445
	FlashVSR	23.840	0.634	0.407	0.985	0.748	0.541
	PS-SR	22.368	0.597	0.326	0.987	0.719	0.497
	SparkVSR	22.962	0.621	0.348	0.984	0.735	0.532
	OSDEnhancer	22.657	0.596	0.307	0.982	0.763	0.522
	Base Model	22.013	0.536	0.375	0.962	0.750	0.519
	FGPO (Ours)	23.878	0.651	0.295	0.989	0.772	0.548
VideoLQ-50	SeedVR2	–	–	–	0.905	0.671	0.519
	FlashVSR	–	–	–	0.896	0.693	0.475
	PS-SR	–	–	–	0.911	0.678	0.509
	SparkVSR	–	–	–	0.906	0.724	0.515
	OSDEnhancer	–	–	–	0.905	0.725	0.506
	Base Model	–	–	–	0.914	0.737	0.570
	FGPO (Ours)	–	–	–	0.925	0.755	0.595

Table 2: Comparison of different training strategies for video super-resolution on the YouHQ-40 dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Training Strategy	PSNR ↑	SSIM ↑	LPIPS ↓	VBench-SC ↑	VBench-IQ ↑	VBench-AQ ↑
Base Model	22.255	0.594	0.364	0.970	0.732	0.565
Standard SFT	<u>23.250</u>	<u>0.635</u>	<u>0.320</u>	<u>0.975</u>	0.736	0.570
Vanilla GRPO	22.780	0.616	0.342	0.972	0.742	0.572
Vanilla DPO	22.950	0.623	0.333	0.973	0.746	0.575
Global Scalar Reward	22.480	0.605	0.352	0.971	0.748	0.576
Global Diversity Regularization	22.560	0.609	0.348	0.971	<u>0.756</u>	<u>0.584</u>
FGPO (Ours)	**24.562**	**0.702**	**0.275**	**0.979**	**0.762**	**0.592**

4.3 Qualitative Results

As shown in Fig. 3, we present qualitative comparisons to evaluate visual realism and structural alignment. In the outdoor pedestrian scene, existing methods generally preserve coarse body structures but suffer from a trade-off between over-smoothing and structural distortion. SeedVR2 (Wang et al., 2026a), FlashVSR (Zhuang et al., 2026), PS-SR (Wu et al., 2026), and the Base Model produce blurry facial regions due to over-regularization, whereas SparkVSR (Yu et al., 2026) and OSDEnhancer (Wei et al., 2026) introduce noticeable facial distortions and artifact-laden garment textures. In contrast, FGPO reconstructs sharper facial features and finer clothing details while maintaining global structural alignment with the input, enhancing high-dispersion local regions with fewer visible alterations to coarse geometry. In the yellow bird scene, existing methods similarly struggle with fine organic textures, either blurring the eye and beak or merging neighboring feather structures into smoothed patches. FGPO, however, preserves overall anatomical structures while recovering distinct eye contours, sharp beak boundaries, and well-separated feather patterns. These visual comparisons suggest that FGPO balances over-smoothing and unconstrained hallucination by selectively applying feature-discrepancy rewards to high-dispersion areas. Complete video results are provided in the supplementary material to illustrate the temporal consistency of the reconstructed results.

Table 3: Ablation study of different components of FGPO on the YouHQ-40 dataset. The best and second-best results are highlighted in **bold** and underlined, respectively.

Feasibility-Gated Partitioning	R_i^{L}	Mask Mechanism	PSNR ↑	SSIM ↑	LPIPS ↓	VBench-SC ↑	VBench-IQ ↑	VBench-AQ ↑
×	×	×	22.780	0.616	0.342	0.972	0.742	0.572
✓	×	×	23.350	0.642	0.320	0.974	0.748	0.578
✓	✓	×	23.750	0.665	0.305	0.975	0.752	0.582
✓	✓	✓ (Unguided)	24.100	0.680	0.288	0.976	0.757	0.586
✓	✓	✓ **(LR-Guided)**	**24.562**	**0.702**	**0.275**	**0.979**	**0.762**	**0.592**



Figure 4: **Visual comparison of different ablation configurations.** The complete FGPO configuration shows a favorable observed trade-off between structural fidelity and detail synthesis.

4.4 Ablation Experiments

As shown in Table 3 and Fig. 4, we conduct an ablation study on YouHQ-40 to evaluate our components. Evaluating feasibility-gated partitioning against Vanilla GRPO confirms its role in restricting optimization eligibility for candidate videos that fail the learned LR-conditioned detail plausibility and temporal consistency criteria. Incorporating this candidate-level feasibility gate markedly boosts structural fidelity and temporal consistency. By assigning lower eligibility to implausible outputs, the gate reduces the risk of reinforcing structurally inconsistent candidate videos, maintaining frame-to-frame stability without forcing ungrounded high-frequency details. Introducing the local feature discrepancy reward R_i^{L} without spatial masking enhances perceptual sharpness and detail synthesis. However, without spatial guidance, applying this reward uniformly across the image fails to separate genuinely ambiguous regions from deterministic structures already constrained by the LR input, potentially introducing subtle artifacts in otherwise stable regions.

Applying an unguided candidate-variance mask mitigates this uniform penalty by restricting R_i^{L} to areas exhibiting high inter-candidate dispersion. This spatial selectivity yields simultaneous improvements across fidelity, temporal consistency, and perceptual metrics over the unmasked variant. However, high variance often appears near sharp edges or motion boundaries simply due to sampling noise rather than underlying visual ambiguity, making raw variance an imperfect spatial criterion. The complete FGPO framework incorporates the LR-guided mask, leveraging LR structural references to refine spatial weighting and reduce edge-associated artifacts. The full model outperforms the unguided-mask variant across all metrics, indicating that LR guidance provides a more effective spatial organization for the reward field. By encouraging local feature variation in high-dispersion regions while assigning lower reward weights around LR-aligned structures, FGPO achieves a favorable balance between detail richness and temporal stability. The visual comparisons are consistent with the complementary roles of each proposed component.

5 Conclusion

In this work, we propose **FGPO**, a Feasibility-Gated Preference Optimization framework for video super-resolution. Unlike existing generative and preference-optimization methods that apply spatially uniform perceptual rewards or feature regularization, FGPO differentiates LR-constrained structures from regions with higher empirical dispersion. Specifically, FGPO partitions sampled candidate videos via learned detail-plausibility and temporal-consistency criteria, uses Pareto analysis over perceptual objectives to determine eligibility for feature-discrepancy rewards, and spatially weights this bonus using an LR-guided candidate-dispersion mask. The resulting feedback can be exposed through sample-level scalar rewards or pairwise preferences, separating its construction from downstream policy update rules. Across the evaluated VSR benchmarks, FGPO consistently elevates its Base Model and achieves a favorable balance among reconstruction fidelity, temporal consistency, and perceptual quality.

REFERENCES

Haoran Bai, Xiaoxu Chen, Canqian Yang, Zongyao He, Sibin Deng, and Ying Chen. Vivid-vr: Distilling concepts from text-to-video diffusion transformer for photorealistic video restoration. *arXiv preprint arXiv:2508.14483*, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237. IEEE, 2018.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3086–3095. IEEE, 2019.

Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4945–4954. IEEE, 2021.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5962–5971. IEEE, 2022a.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5952–5961. IEEE, 2022b.

Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. Dove: Efficient one-step diffusion model for real-world video super-resolution. *Advances in Neural Information Processing Systems*, 38:85218–85237, 2026.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4):75–1, 2020.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, et al. Sensenova-u1. 5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Google DeepMind. Gemini 3.8 flash model card, 2026. URL `https://deepmind.google/models/model-cards/gemini-3-8-flash/`.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3224–3232. IEEE, 2018.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 105–114. IEEE, 2017.

Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in neural information processing systems*, 35:378–393, 2022.

Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 33:2171–2182, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2026b.

Yidi Liu, Zihao Fan, Jie Huang, Jie Xiao, Dong Li, Wenlong Zhang, Lei Bai, Xueyang Fu, and Zheng-Jun Zha. Finpercep-rm: A fine-grained reward model and co-evolutionary curriculum for rl-based real-world super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4839–4849, 2026c.

Seungjun Nah, Radu Timofte, Shuhang Gu, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Kyoung Mu Lee, Xintao Wang, Kelvin CK Chan, et al. Ntire 2019 challenge on video super-resolution: Methods and results. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1985–1995. IEEE, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Charles Poynton. *Digital video and HD: Algorithms and Interfaces*. Elsevier, 2012.

Junbo Qiao, Miaomiao Cai, Wei Li, Xudong Huang, Jie Hu, Xinghao Chen, Shaohui Lin, and Hongkai Xiong. Realsr-r1: Reinforcement learning for real-world image super-resolution with vision-language chain-of-thought. *arXiv preprint arXiv:2506.16796*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *2017 IEEE Conference on computer vision and pattern recognition (CVPR)*, pp. 237–246. IEEE, 2017.

Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4482–4490. IEEE, 2017.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, et al. Seedvr2: One-step video restoration via diffusion adversarial post-training. In *International Conference on Learning Representations*, volume 2026, pp. 40987–41009, 2026a.

MengChao Wang, Wang Qiang, Fan Jiang, and Mu Xu. Fantasytalking2: Timestep-layer adaptive preference optimization for audio-driven portrait animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9966–9974, 2026b.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *European conference on computer vision*, pp. 63–79. Springer, 2018.

Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1954–1963. IEEE, 2019.

Shuoyan Wei, Feng Li, Chen Zhou, Runmin Cong, Yao Zhao, and Huihui Bai. Taming real-world space-time video super-resolution with one-step diffusion. *arXiv preprint arXiv:2601.20308*, 2026.

Aiqiu Wu, Zhaofan Qiu, Ting Yao, and Tao Mei. Ps-sr: Pseudo-single-step video super-resolution via speculative diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38218–38227, 2026.

Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17108–17118. IEEE, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Yiran Xu, Taesung Park, Richard Zhang, Yang Zhou, Eli Shechtman, Feng Liu, Jia-Bin Huang, and Difan Liu. Videogigagan: Towards detail-rich video super-resolution. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2139–2149. IEEE, 2025.

Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4761–4770. IEEE, 2021.

Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3106–3115. IEEE, 2019.

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, and Lei Zhang. Gdpo-sr: Group direct preference optimization for one-step generative image super-resolution. *arXiv preprint arXiv:2603.16769*, 2026.

Jiongze Yu, Xiangbo Gao, Pooja Verlani, Akshay Gadde, Yilin Wang, Balu Adsumilli, and Zhengzhong Tu. Sparkvsr: Interactive video super-resolution via sparse keyframe propagation. In *European Conference on Computer Vision*, pp. 397–415. Springer, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2535–2545. IEEE, 2024.

Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. Flashvsr: Towards real-time diffusion-based streaming video super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43482–43493, 2026.

A AI ASSISTANCE STATEMENT

During manuscript preparation, we used an artificial intelligence (AI) language model to assist with writing and language polishing, including grammar checking, sentence rephrasing, readability improvement, and clarity enhancement. The AI tool was not involved in the research ideas, methodology, experimental design, data collection, or data analysis; all scientific content and decisions were developed and verified by the authors. We ensured that the AI-generated or polished text complies with ethical guidelines and does not contribute to plagiarism or scientific misconduct.

B CONSTRUCTION AND VALIDATION OF THE MULTI-OBJECTIVE VIDEO EVALUATOR

This section describes the supervised fine-tuning of our locally deployed Qwen3.5-Omni evaluator, which provides task-specific, low-latency online feedback during preference optimization. We detail the multi-objective evaluation framework, the targeted prompt design, dataset construction, parameter-efficient fine-tuning configurations, and the validation of the fine-tuned evaluator.

B.1 HUMAN ANNOTATION PROTOCOL AND DATASET PARTITION

To establish a rigorous, human-aligned ground truth for training and evaluating our multimodal evaluator, we recruited 30 experienced human annotators with professional backgrounds in computer vision, video editing, and digital media processing. Prior to annotation, all participants underwent standardized training to familiarize themselves with the four evaluation dimensions (q^D, q^T, q^C, and q^A) and the 1-to-10 scoring criteria.

Inter-Annotator Agreement and Discrepancy Resolution. Each video sample was independently rated by multiple annotators under controlled viewing conditions. To quantify inter-annotator consistency, we computed Krippendorff's α across all scoring dimensions, achieving a high overall agreement of $\alpha = 0.842$ ($\alpha > 0.80$ indicates strong reliability). For samples exhibiting score variance exceeding 1.5 points, a consensus adjudication procedure was triggered: the divergent ratings were re-inspected in joint panel discussions led by two senior visual quality experts to reconcile discrepancies and assign a finalized, consensus-backed score.

Evaluator Training and Validation Splitting. The complete annotated dataset comprises $20,000$ dimension-specific sample triplets. To ensure robust generalization and prevent evaluation leakage during post-training, the video sequences were partitioned into a training set of $16,000$ samples (80%) and a validation set of $4,000$ samples (20%). The split was conducted strictly at the video sequence level, ensuring that no overlapping scenes or source videos appear in both training and validation splits.

B.2 UNIFIED MULTI-OBJECTIVE EVALUATION PARADIGM

In our framework, the multidimensional evaluation matrix $\mathbf{Q} \in \mathbb{R}^{K\times 4}$ comprises four distinct evaluation dimensions: detail plausibility q^D, temporal consistency q^T, spatial clarity q^C, and aesthetic quality q^A. The fine-tuned Qwen3.5-Omni evaluator is trained to assess these dimensions using dimension-specific structured evaluation instructions.

For each candidate video, the evaluator performs forward inference to generate a structured evaluation rationale along with an integer quality rating $s^d \in \{1, \ldots, 10\}$. The score is then normalized to a scale of $[0, 1]$ via:

$$q^d = \frac{s^d - 1}{9}, \quad d \in \{D, T, C, A\}. \tag{7}$$

These normalized scores populate the matrix $\mathbf{Q}$ in real-time, enabling our progressively gated reinforcement learning algorithm to execute spatio-temporally differentiated policy updates without relying on external online APIs during reinforcement learning.

`Qwen3.5-Omni` Targeted Quality Evaluation Prompt Template

[System Role]
You are an expert visual quality evaluator specializing in video super-resolution (VSR) assessment, structural fidelity, and temporal stability analysis.

[Input Data]
- High-Resolution (HR) Video Sequence: $\mathbf{x}_i$ (the generated candidate to be evaluated)
- Low-Resolution (LR) Reference Video: $\mathbf{y}$ (the ground-truth input guiding the restoration)

[Active Evaluation Objective]
You must evaluate the video sequence strictly along the target dimension: **[Target Dimension]**
(The prompt dynamically injects one of the following four criteria based on the active sub-task:)
- **Detail Plausibility** (q^D): Assess how faithfully the HR video preserves the structural geometry, text characters, and object identities supported by the LR reference $\mathbf{y}$. Penalize ungrounded hallucinations (e.g., distorted text, modified facial identities, or fabricated geometric patterns).
- **Temporal Consistency** (q^T): Assess the frame-to-frame consistency and motion stability of the HR sequence. Penalize temporal defects such as inter-frame flickering, sudden structural warping, or "texture swimming" on static surfaces.
- **Spatial Clarity** (q^C): Assess the sharpness of high-frequency details, edge definitions, and microscopic texture recovery. Penalize blurred regions, muddy details, or residual upscaling artifacts.
- **Aesthetic Quality** (q^A): Assess the overall visual realism, color naturalness, and perceptual appeal of the HR video. Penalize unnatural artifacts such as severe over-sharpening halos, plastic-like over-smoothing, or artificial color casts.

[Structured Evaluation Protocol]
To guarantee objective, fine-grained, and logical scoring, you must perform a structured analysis covering:
- *Spatio-Temporal Evidence*: Compare the LR and HR videos to identify successfully restored texture regions and remaining distorted zones.
- *Defect Localization*: Point out specific artifacts corresponding strictly to the active target dimension.
- *Score Justification*: Summarize your findings to justify the final score for this target dimension.

[Output Format]
Provide your structured evaluation rationale first, then output the final single integer score (ranging from 1 to 10) on the very last line strictly using the following format for programmatic parsing:
Rationale: [Your detailed structured quality analysis text]
Score: [Active Dimension: X] (where X is an integer from 1 to 10 corresponding to the active dimension)

Figure 5: The targeted evaluation prompt template for dimension-specific quality assessment.

B.3 Multi-Objective Prompt Design and Dataset Construction

To distill the advanced perceptual capabilities of Gemini-3.8-Flash into our local Qwen3.5-Omni, we construct an instruction-tuning dataset based on the RealVSR dataset (Yang et al., 2021). We perform multiple dimension-specific evaluations on the videos: each video is evaluated multiple times, with each evaluation run targeting exactly one of the four quality dimensions using the template in Fig. 5. This process yields exactly $5,000$ annotated samples for each of the four dimensions (q^D, q^T, q^C, and q^A), resulting in a perfectly balanced dataset of $\mathcal{S} = \{(x_i, I_i, T_i)\}_{i=1}^{20,000}$ triplets. The target text T_i consists of a structured evaluation rationale followed by a single dimension-specific score (1 to 10), formatted for direct programmatic parsing.

B.4 Parameter-Efficient Fine-Tuning and Hyperparameters

We perform autoregressive supervised fine-tuning on the Base Model Qwen3.5-Omni. To optimize GPU memory and preserve its general multimodal interaction capabilities, we employ Low-Rank Adaptation (LoRA), where the trainable low-rank weight update is defined as $\Delta\mathbf{W} = \frac{\alpha}{r}\mathbf{BA}$, with rank $r = 16$ and scaling factor $\alpha = 32$. The training objective minimizes the standard autoregressive cross-entropy loss over the target tokens:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{\sum_{i=1}^{M} N_i} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \log p_\theta(t_{i,j} \mid t_{i,<j}, x_i, I_i) \tag{8}$$

Table 4: Quantitative evaluation agreement analysis of Qwen3.5-Omni before and after SFT across four evaluation dimensions.

	Against Teacher Model (Gemini-3.8-Flash)			Against Human Consensus		
Evaluation Dimension and Config	**PCC** ↑	**SRCC** ↑	**Acc$_{\pm1}$** ↑	**PCC** ↑	**SRCC** ↑	**Acc$_{\pm1}$** ↑
Detail Plausibility Dimension (q^D)						
Vanilla Qwen3.5-Omni	0.6820	0.6515	63.4%	0.6215	0.5902	58.5%
Fine-tuned Qwen3.5-Omni (Ours)	**0.9782**	**0.9680**	**95.2%**	**0.9610**	**0.9520**	**92.1%**
Temporal Consistency Dimension (q^T)						
Vanilla Qwen3.5-Omni	0.6505	0.6292	60.1%	0.5952	0.5720	55.4%
Fine-tuned Qwen3.5-Omni (Ours)	**0.9715**	**0.9580**	**94.0%**	**0.9520**	**0.9410**	**90.5%**
Spatial Clarity Dimension (q^C)						
Vanilla Qwen3.5-Omni	0.7152	0.6812	65.8%	0.6420	0.6150	60.2%
Fine-tuned Qwen3.5-Omni (Ours)	**0.9750**	**0.9620**	**94.8%**	**0.9580**	**0.9480**	**91.8%**
Aesthetic Quality Dimension (q^A)						
Vanilla Qwen3.5-Omni	0.7320	0.7005	68.2%	0.6650	0.6380	62.5%
Fine-tuned Qwen3.5-Omni (Ours)	**0.9680**	**0.9510**	**93.2%**	**0.9490**	**0.9380**	**89.8%**

where θ denotes the trainable parameters, $t_{i,j}$ is the j-th token of the target rationale and score text T_i, N_i represents the token length of T_i, and $M = 20,000$ is the total dataset size.

For the supervised fine-tuning process, we utilize the AdamW optimizer with a weight decay coefficient of 0.01. The training is executed on an 8-node NVIDIA H800 GPU cluster with a global batch size of 32. The initial learning rate is set to 2×10^{-5} and scheduled using a cosine annealing decay strategy with a warmup phase covering the first 10% of the total steps. The entire training process is conducted for $20,000$ optimization steps , ensuring stable alignment without over-fitting to the teacher's stylistic templates.

B.5 Quantitative Validation and Agreement Analysis

To evaluate the reliability of the fine-tuned Qwen3.5-Omni(Ours) in assessing detailed structural constraints and perceptual quality, we evaluate on UDM10 (Yi et al., 2019) comprising unseen VSR test videos. We measure the alignment of the evaluator before and after fine-tuning against two reference standards: the teacher model (Gemini-3.8-Flash) scores, and the average ratings from human visual experts (Human Consensus).

We adopt three standard statistical metrics: (1) Pearson Correlation Coefficient (PCC), evaluating the linear correlation between predicted and reference ratings; (2) Spearman's Rank Correlation Coefficient (SRCC), evaluating monotonic rank agreement, which is crucial for preference pair construction; (3) Thresholded Tolerance Accuracy (Acc$_{\pm1}$), measuring the percentage of samples where the absolute error of the predicted rating is within ±1 of the reference score.

The quantitative results are reported in Table 4. First, the fine-tuned model illustrates an excellent understanding of detailed structural constraints. In the detail plausibility (q^D) and temporal consistency (q^T) dimensions, the vanilla Qwen3.5-Omni exhibits moderate but suboptimal detailed structural alignment with the reference ratings, achieving SRCC values of only 0.6515 and 0.6292 against the teacher, and 0.5902 and 0.5720 against human consensus, respectively. In contrast, supervised fine-tuning substantially increases the corresponding SRCC values against the teacher model to 0.9680 and 0.9580. These improvements suggest that the aligned evaluator exhibits increased sensitivity toward fine geometric misalignments, structural distortions, and high-frequency temporal flickering across adjacent frames. Moreover, the aligned Qwen3.5-Omni exhibits consistent prediction stability, achieving a thresholded tolerance accuracy (Acc$_{\pm1}$) above 90.0% across all four evaluation dimensions. Such consistent scoring helps maintain stable reward magnitude distributions, mitigating potential policy fluctuations and scale drift during post-training. Finally, when evaluated against Human Consensus, the aligned model achieves near-perfect agreement across all objectives, with SRCC values consistently reaching around 0.95 (exceeding 0.94 across all dimensions). This result indicates that the locally deployed model effectively captures human visual preferences regarding structural plausibility and temporal consistency, as well as high-level perceptual aesthetics, thereby providing reliable and human-aligned feedback signals for subsequent policy optimization.

C FGPO WITH ALTERNATIVE PREFERENCE-OPTIMIZATION OBJECTIVES

To validate the generalizability of FGPO beyond policy-gradient RL (e.g., GRPO), we instantiate FGPO under offline pairwise preference optimization via Diffusion DPO (Wallace et al., 2024) and direct reward-based Flow Matching optimization via DiffusionNFT (Zheng et al., 2026).

C.1 TWO-TIER MARGIN-AWARE PREFERENCE PAIR CONSTRUCTION FOR DPO

To convert FGPO's multi-objective feedback matrix $\mathbf{Q} \in \mathbb{R}^{K\times 4}$ and local feature-discrepancy rewards into explicit pairwise preference supervision $\mathcal{I}_{\text{pair}} = \{(x_w^{(m)}, x_l^{(m)})\}_{m=1}^{M}$ for DPO without introducing ambiguous signals or self-contradictions, we establish a deterministic, two-tier margin-aware pairing protocol, where $x_w^{(m)}$ and $x_l^{(m)}$ denote the preferred (winner) and dispreferred (loser) candidate videos in the m-th pair, and M is the total count of constructed preference pairs.

Cross-Partition Feasibility Pair Construction. To strictly enforce physical feasibility boundaries during offline preference learning, any gate-passing feasible candidate $x_w \in \mathcal{F}$ is assigned higher preference over any infeasible candidate $x_l \in \mathcal{D}$, denoted as $x_w \succ x_l$ where $\succ$ represents the strict preference relation. Infeasible candidates in $\mathcal{D}$ are ordered by their continuous violation penalty R^{V}. To eliminate noisy pairs caused by evaluator jitter, a cross-partition pair is retained only if the scalar feedback gap satisfies $\widetilde{R}_w - \widetilde{R}_l > \gamma_{\text{margin}}$, where $\widetilde{R}_w$ and $\widetilde{R}_l$ are the scalar feedback values (Sec. 3.4) and $\gamma_{\text{margin}} \geq 0$ is a scalar safety margin hyperparameter that filters out marginal feedback fluctuations; exact score ties ($\widetilde{R}_w = \widetilde{R}_l$) are automatically discarded.

Intra-Feasible Discrepancy-Guided Pair Construction. Within the feasible candidate set $\mathcal{F}$, candidates are ranked directly by their unified scalar feedback $\widetilde{R}_i$. Because candidates in the first-front Pareto set $\mathcal{P}$ receive the localized feature-discrepancy bonus R_i^{L}, they naturally attain higher scalar feedback when exhibiting plausible structural variations. If a non-front candidate $x_k \in \mathcal{F} \setminus \mathcal{P}$ possesses an exceptionally high base quality score R_k^{Q} that surpasses a front candidate's total positive reward $R_j^{\text{P}} = R_j^{\text{Q}} + R_j^{\text{L}}$, the explicit preference directly follows $\widetilde{R}_k > \widetilde{R}_j$, ensuring complete mathematical alignment between DPO pairwise preference ordering and GRPO advantage estimation. Intra-feasible pairs are sampled uniformly from all valid combinations satisfying $\widetilde{R}_w - \widetilde{R}_l > \gamma_{\text{margin}}$, capping the maximum cohort pair capacity at $M \leq \binom{|\mathcal{F}|}{2}$, where $\binom{|\mathcal{F}|}{2} = \frac{|\mathcal{F}|(|\mathcal{F}|-1)}{2}$ denotes the combination number of choosing two candidates from the feasible set size $|\mathcal{F}|$.

C.2 QUANTITATIVE ANALYSIS ACROSS OPTIMIZATION OBJECTIVES

As reported in Tables 5 and 6, both the Diffusion DPO (Wallace et al., 2024) and DiffusionNFT (Zheng et al., 2026) variants of FGPO outperform all existing VSR baselines and achieve substantial improvements over the Base Model across all evaluated benchmarks, highlighting the framework's versatility. Meanwhile, the two variants exhibit complementary strengths in the trade-off between reconstruction fidelity and perceptual quality. The DPO variant favors perceptual sharpness and aesthetic quality, achieving optimal performance in LPIPS (Zhang et al., 2018), VBench-IQ, and VBench-AQ (Huang et al., 2024). In contrast, the DiffusionNFT variant emphasizes reconstruction fidelity, obtaining higher PSNR and SSIM (Zhang et al., 2018) as well as superior subject consistency over time (VBench-SC (Huang et al., 2024)) while maintaining top-tier perceptual quality. These results demonstrate that FGPO's feasibility-gated Pareto-front design remains robust and effective across diverse preference-optimization objectives.

D VISUALIZATION OF DISPERSION-BASED WEIGHTING MASKS AND ADAPTIVE WEIGHT FIELDS

Fig. 6 visualizes the construction of the LR-guided spatiotemporal dispersion-based weighting mask and the resulting adaptive weight field. Given multiple candidate videos generated from the same LR input, we convert their frames to the YCbCr space and extract the luminance channels. The

Table 5: Quantitative comparison on four video super-resolution benchmarks using FGPO (Ours with DPO). The best and second-best results are highlighted in red and orange, respectively.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	VBench-SC ↑	VBench-IQ ↑	VBench-AQ ↑
YouHQ-40	SeedVR2	21.167	0.632	0.376	0.948	0.630	0.526
	FlashVSR	23.130	0.604	0.387	0.975	0.745	0.580
	PS-SR	23.893	0.669	0.312	0.978	0.730	0.569
	SparkVSR	24.443	0.686	0.343	0.975	0.725	0.568
	OSDEnhancer	24.322	0.676	0.289	0.974	0.735	0.548
	Base Model	22.255	0.594	0.364	0.970	0.732	0.565
	FGPO (Ours with DPO)	24.525	0.698	0.278	0.981	0.768	0.598
REDS-30	SeedVR2	22.192	0.599	0.429	0.839	0.526	0.461
	FlashVSR	23.920	0.649	0.419	0.896	0.741	0.515
	PS-SR	23.552	0.621	0.324	0.901	0.679	0.475
	SparkVSR	23.196	0.594	0.454	0.900	0.616	0.483
	OSDEnhancer	22.936	0.587	0.334	0.896	0.714	0.473
	Base Model	21.549	0.535	0.365	0.890	0.751	0.504
	FGPO (Ours with DPO)	24.015	0.658	0.312	0.905	0.772	0.542
SPMCS-30	SeedVR2	19.735	0.539	0.447	0.887	0.502	0.445
	FlashVSR	23.840	0.634	0.407	0.985	0.748	0.541
	PS-SR	22.368	0.597	0.326	0.987	0.719	0.497
	SparkVSR	22.962	0.621	0.348	0.984	0.735	0.532
	OSDEnhancer	22.657	0.596	0.307	0.982	0.763	0.522
	Base Model	22.013	0.536	0.375	0.962	0.750	0.519
	FGPO (Ours with DPO)	23.865	0.645	0.292	0.989	0.775	0.555
VideoLQ-50	SeedVR2	–	–	–	0.905	0.671	0.519
	FlashVSR	–	–	–	0.896	0.693	0.475
	PS-SR	–	–	–	0.911	0.678	0.509
	SparkVSR	–	–	–	0.906	0.724	0.515
	OSDEnhancer	–	–	–	0.905	0.725	0.506
	Base Model	–	–	–	0.914	0.737	0.570
	FGPO (Ours with DPO)	–	–	–	0.928	0.762	0.601

Table 6: Quantitative comparison on four video super-resolution benchmarks using FGPO (Ours with DiffusionNFT). The best and second-best results are highlighted in red and orange, respectively.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	VBench-SC ↑	VBench-IQ ↑	VBench-AQ ↑
YouHQ-40	SeedVR2	21.167	0.632	0.376	0.948	0.630	0.526
	FlashVSR	23.130	0.604	0.387	0.975	0.745	0.580
	PS-SR	23.893	0.669	0.312	0.978	0.730	0.569
	SparkVSR	24.443	0.686	0.343	0.975	0.725	0.568
	OSDEnhancer	24.322	0.676	0.289	0.974	0.735	0.548
	Base Model	22.255	0.594	0.364	0.970	0.732	0.565
	FGPO (Ours with DiffusionNFT)	24.545	0.701	0.284	0.983	0.755	0.586
REDS-30	SeedVR2	22.192	0.599	0.429	0.839	0.526	0.461
	FlashVSR	23.920	0.649	0.419	0.896	0.741	0.515
	PS-SR	23.552	0.621	0.324	0.901	0.679	0.475
	SparkVSR	23.196	0.594	0.454	0.900	0.616	0.483
	OSDEnhancer	22.936	0.587	0.334	0.896	0.714	0.473
	Base Model	21.549	0.535	0.365	0.890	0.751	0.504
	FGPO (Ours with DiffusionNFT)	24.025	0.661	0.318	0.907	0.760	0.529
SPMCS-30	SeedVR2	19.735	0.539	0.447	0.887	0.502	0.445
	FlashVSR	23.840	0.634	0.407	0.985	0.748	0.541
	PS-SR	22.368	0.597	0.326	0.987	0.719	0.497
	SparkVSR	22.962	0.621	0.348	0.984	0.735	0.532
	OSDEnhancer	22.657	0.596	0.307	0.982	0.763	0.522
	Base Model	22.013	0.536	0.375	0.962	0.750	0.519
	FGPO (Ours with DiffusionNFT)	23.872	0.648	0.297	0.991	0.768	0.549
VideoLQ-50	SeedVR2	–	–	–	0.905	0.671	0.519
	FlashVSR	–	–	–	0.896	0.693	0.475
	PS-SR	–	–	–	0.911	0.678	0.509
	SparkVSR	–	–	–	0.906	0.724	0.515
	OSDEnhancer	–	–	–	0.905	0.725	0.506
	Base Model	–	–	–	0.914	0.737	0.570
	FGPO (Ours with DiffusionNFT)	–	–	–	0.929	0.750	0.585

luminance variations across candidates yield the empirical dispersion field, capturing the spatial and temporal disagreement induced by stochastic generation. Because raw dispersion may contain noisy, fragmented responses near local edges, we apply the spatiotemporal guided filter (ST-GF) using the

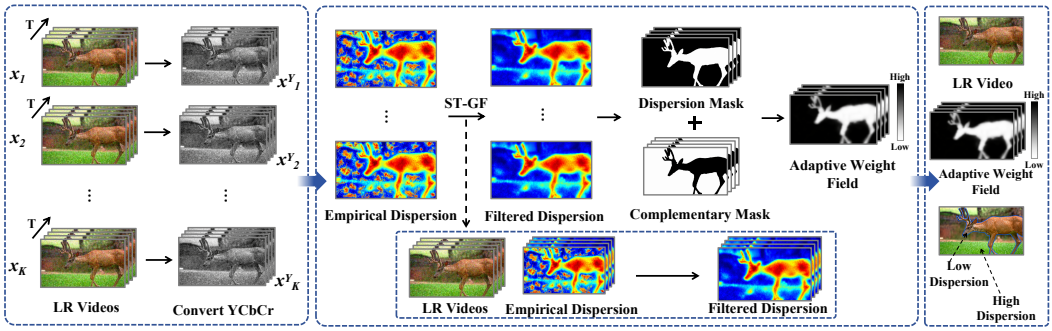

Figure 6: Visualization of LR-guided spatiotemporal empirical candidate-dispersion estimation and adaptive weight-field construction.

Table 7: **Hyperparameter Specifications for Core FGPO Components.**

Hyperparameter / Setting	**Value**
Detail Plausibility Weight (w_D)	1.0
Temporal Consistency Weight (w_T)	1.0
Spatial Clarity Weight (w_C)	0.5
Aesthetic Quality Weight (w_A)	0.5
Detail Plausibility Threshold (τ_D)	0.15
Temporal Consistency Threshold (τ_T)	0.85
Gate Steepness (β_D, β_T)	10.0, 10.0
Violation Normalization (s_D, s_T)	0.1, 0.1
Single Violation Weights (λ_D, λ_T)	1.0, 1.0
Joint Violation Weight (λ_{DT})	0.5
Dispersion Bounds ($w_{\text{det}}, w_{\text{stoch}}$)	0.1, 1.0

upsampled LR luminance as guidance. The filtered dispersion exhibits smoother, structure-aligned responses while suppressing leakage across LR-observable boundaries.

The filtered dispersion is subsequently normalized and mapped to the dispersion-based weighting mask M. Its complementary mask $1 - M$ highlights relatively stable regions where LR constraints are strong, whereas high-response regions in M indicate high-dispersion areas where local discrepancies should be evaluated. The two masks are combined to produce the spatiotemporally adaptive weight field W_{F}. As illustrated on the right, the resulting field assigns lower weights to low-dispersion structures and higher weights to high-dispersion regions. Thus, unlike candidate-level feasibility gating—which partitions whole candidates—the mask construction imposes no hard pixel-level partition; instead, it provides a soft spatiotemporal weighting signal that moderates feature-discrepancy rewards around LR-supported structures while encouraging feature-space exploration in empirically high-dispersion regions.

E Hyperparameter Specifications

To ensure exact reproducibility, we provide the complete list of hyperparameter settings and algorithmic parameters used in FGPO in Table 7.

F Human Study Results

To evaluate the perceptual quality of reconstructed videos beyond conventional quantitative metrics, we conduct a human preference study involving 20 experienced assessors under a randomized, double-blind setup. The evaluation dataset comprises 30 video sequences randomly sampled from multiple real-world and high-quality benchmarks such as YouHQ-40 and SPMCS-30, covering

Table 8: **Human Study Results.** Mean Opinion Scores (1–5 scale) on Visual Realism (VR), Temporal Smoothness (TS), and Overall Quality (OQ). Among all evaluated restoration methods (excluding Ground Truth), the best and second-best results are highlighted in red and orange, respectively.

Method	Visual Realism (VR) ↑	Temporal Smoothness (TS) ↑	Overall Quality (OQ) ↑
Ground Truth (GT)	*4.82*	*4.88*	*4.85*
SeedVR2	2.85	3.10	2.92
FlashVSR	3.42	3.85	3.58
PS-SR	3.65	3.40	3.55
SparkVSR	3.50	3.52	3.51
OSDEnhancer	3.88	3.60	3.75
Base Model	3.12	3.70	3.42
FGPO (Ours)	4.35	4.28	4.32

complex human motions, rich visual textures, and diverse degradation patterns. Participants independently assess the generated videos across three perceptual dimensions: Visual Realism (VR), Temporal Smoothness (TS), and Overall Quality (OQ). Ratings are collected using a 5-point Mean Opinion Score (MOS) scale, with explicitly defined scoring anchors: *1 = Very Poor (severe artifacts/unacceptable quality)*, *3 = Fair (acceptable quality with minor flaws)*, and *5 = Excellent (imperceptible artifacts/real-world realism)*.

As summarized in Table 8, among all evaluated VSR and generative restoration methods, FGPO consistently achieves the highest MOS scores across all three dimensions, outperforming existing state-of-the-art baselines. Furthermore, the human ratings for FGPO are the closest to the Ground Truth videos, which serve as the reference upper bound. These results demonstrate that our preference optimization framework aligns generated videos more effectively with human visual perception and real-world distributions.

G Quantitative Analysis of Dispersion Normalization Strategies

To evaluate the mathematical behavior of candidate-dispersion normalization, we quantitatively benchmark four normalization formulations across synthetic REDS-30 (Nah et al., 2019) and real-world VideoLQ-50 (Chan et al., 2022b) validation sequences: raw unnormalized variance ($\widetilde{V}$), percentile mapping (M_{perc}), dataset-level calibrated gating (M_{ds}), and our per-video mean normalization (M). As reported in Table 9, unnormalized variance fluctuates drastically across domain shifts, which would induce severe scale explosion or vanishing in policy-gradient estimations if directly applied to reward calculations. Dataset-level global calibration suffers from severe domain imbalance by under-activating on clean sequences (22.1%) while over-activating on complex sequences (78.9%), whereas percentile mapping overly compresses intra-video spatial contrast ($\sigma_M \approx 0.21$). In contrast, our per-video mean normalization (M) maintains a balanced exploration coverage ($38.5\% \sim 41.2\%$) across diverse content while preserving high spatial contrast ($\sigma_M \geq 0.31$), providing stable reward-scale calibration and effective localized feature-discrepancy guidance during reinforcement learning.

H Calibration and Pass-Rate Analysis of Feasibility Thresholds

The asymmetric selection of $\tau_D = 0.15$ and $\tau_T = 0.85$ directly addresses the physical failure modes inherent to video super-resolution. Given the evaluator's linear mapping $q = (s-1)/9$ from raw scores $s \in [1, 10]$, temporal flickering represents a catastrophic failure in video generation, warranting a strict threshold ($\tau_T = 0.85$, corresponding to an absolute evaluator score $s_T \geq 8.65/10$) to immediately penalize unstable candidates. Conversely, local detail artifacts require a calibrated safety threshold ($\tau_D = 0.15$, corresponding to score $s_D \geq 2.35/10$) to filter out severely corrupted or collapsed outputs without prematurely suppressing plausible candidate variations. To verify the operational stability and active filtering contribution of this asymmetric design, we measure empirical candidate pass rates across $10,000$ sampled candidates throughout training. As reported in Table 10, the detail threshold τ_D actively filters out 21.8% of defective candidate samples in the

Table 9: Quantitative comparison of dispersion normalization strategies across datasets. We report the spatial mean of mask responses (μ_M), spatial response standard deviation (σ_M), and the high-exploration pixel coverage ratio ($r_{>0.5}$, percentage of spatiotemporal locations with mask weight > 0.5).

Dataset	Normalization Strategy	Mean (μ_M)	Std (σ_M)	Coverage $r_{>0.5}$ (%)
REDS-30	Raw Unnormalized ($\widetilde{V}$)	1.24×10^{-3}	0.92×10^{-3}	N/A
	Percentile Mapping (M_{perc})	0.48	0.22	32.4%
	Dataset Calibration (M_{ds})	0.39	0.28	22.1%
	Per-Video Mean (Ours, M)	**0.51**	**0.31**	**38.5%**
VideoLQ-50	Raw Unnormalized ($\widetilde{V}$)	3.12×10^{-3}	2.45×10^{-3}	N/A
	Percentile Mapping (M_{perc})	0.49	0.21	32.1%
	Dataset Calibration (M_{ds})	0.76	0.18	78.9%
	Per-Video Mean (Ours, M)	**0.52**	**0.33**	**41.2%**

early training phase, confirming its role as an indispensable safety net. Meanwhile, the feasible set pass rate increases smoothly from 48.9% to 79.6% as training progresses, demonstrating stable and non-collapsing optimization dynamics.

Table 10: Empirical pass rates (%) of candidate cohorts under feasibility gating during FGPO training phases.

Training Phase	Detail ($q^D \geq \tau_D$)	Temporal ($q^T \geq \tau_T$)	Feasible ($\mathcal{F}$)	Pareto Ratio ($\lvert\mathcal{P}\rvert/\lvert\mathcal{F}\rvert$)
Early Phase (Steps 0–500)	78.2%	62.5%	48.9%	31.2%
Mid Phase (Steps 500–1500)	87.5%	75.2%	65.8%	24.5%
Late Phase (Steps 1500–3000)	94.6%	84.2%	79.6%	18.3%

I Dataset Configurations, Baseline Inference Details, and Evaluation Protocols

To ensure complete transparency and rigorous experimental fairness, all methods are evaluated across four representative $4\times$ video super-resolution benchmarks. Specifically, REDS-30 (Nah et al., 2019) comprises the complete 30 validation sequences from the official REDS benchmark (Nah et al., 2019), evaluated using the official low-resolution and high-resolution video pairs. Following the inference configurations of FlashVSR (Zhuang et al., 2026) and Vivid-VR (Bai et al., 2025), SPMCS-30 (Tao et al., 2017) and YouHQ-40 (Zhou et al., 2024) are evaluated using the synthetic degradation pipeline. VideoLQ-50 (Chan et al., 2022b) (50 real-world sequences) features authentic, unknown compound camera noise, compression artifacts, and motion blur, and is evaluated directly without synthetic modification.

Unless otherwise specified, all compared methods are evaluated using publicly available official checkpoints released by their respective authors. A fixed random seed (42) is applied whenever stochastic inference is applicable, and empty prompt strings are supplied for prompt-conditioned models, while models without textual conditioning are evaluated via their standard inference interfaces. No test-time candidate selection, reranking, or Best-of-N filtering is applied during final benchmark evaluation ($N = 1$). Inference step configurations strictly follow each method's official default sampling schedule: SeedVR2 (Wang et al., 2026a), FlashVSR (Zhuang et al., 2026), SparkVSR (Yu et al., 2026), and OSDEnhancer (Wei et al., 2026) rely on single-step forward inference, while PS-SR (Wu et al., 2026) features both single-step forward and lightweight 3-step draft inference. SparkVSR (Yu et al., 2026) is natively an interactive, keyframe-conditioned model; for strict fairness against automated end-to-end VSR methods, we evaluate it under its official blind no-reference (`no_ref`) mode, operating without externally provided or ground-truth reference keyframes. All quantitative metrics are computed using official evaluation implementations in the RGB color space without edge shaving or border cropping, employing official PyTorch implementations for PSNR and SSIM, the official LPIPS (Zhang et al., 2018) package with the AlexNet backbone, and the official VBench evaluation suite for generative video quality metrics.

J Inference Randomness Analysis

To analyze the impact of inference-stage randomness and evaluate performance stability across random seeds, we conduct multi-seed evaluations across 10 identical, randomly selected seeds ($s \in \{10, 20, \ldots, 100\}$) for all compared methods under identical input conditions, model weights, and sampling parameters. Table 11 reports the quantitative performance as Mean $\pm$ Std.

Experimental results show that although our method exhibits a slightly larger variance due to generative feature exploration, its average metrics across 10 random runs remain strictly optimal across all dimensions. The superior VBench-AQ score (0.585 ± 0.008) indicates that our method consistently attains optimal aesthetic quality. This demonstrates that while our approach provides greater generative diversity than state-of-the-art baselines, the generated variations consist of meaningful and plausible detail textures rather than high-frequency noise or meaningless artifacts. This property not only facilitates the reinforcement learning training process by providing diverse yet valid candidate cohorts, but also during inference demonstrates superior expressiveness and flexibility by enabling multiple plausible high-quality visual interpretations under identical low-resolution inputs.

Table 11: Quantitative performance (Mean $\pm$ Std) under multi-seed evaluation (10 identical random seeds) on YouHQ-40.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	VBench-SC ↑	VBench-IQ ↑	VBench-AQ ↑
SeedVR2	21.085 ± 0.12	0.628 ± 0.005	0.381 ± 0.006	0.945 ± 0.004	0.624 ± 0.008	0.521 ± 0.007
FlashVSR	23.214 ± 0.09	0.608 ± 0.004	0.382 ± 0.005	0.973 ± 0.003	0.741 ± 0.006	0.575 ± 0.006
PS-SR	23.815 ± 0.10	0.665 ± 0.004	0.317 ± 0.004	0.976 ± 0.002	0.725 ± 0.006	0.563 ± 0.006
SparkVSR	24.368 ± 0.11	0.681 ± 0.005	0.348 ± 0.006	0.973 ± 0.003	0.719 ± 0.007	0.562 ± 0.007
OSDEnhancer	24.256 ± 0.08	0.672 ± 0.003	0.294 ± 0.004	0.972 ± 0.002	0.731 ± 0.005	0.543 ± 0.005
Base Model	22.180 ± 0.12	0.589 ± 0.005	0.369 ± 0.005	0.967 ± 0.003	0.726 ± 0.007	0.559 ± 0.006
FGPO (Ours)	**24.716** ± 0.15	**0.696** ± 0.006	**0.281** ± 0.005	**0.977** ± 0.003	**0.755** ± 0.009	**0.585** ± 0.008

K Statistical Stability and Ablation of Pareto Selection

Because the task-specific evaluator outputs discrete scores mapped to $[0, 1]$, score ties naturally occur during non-dominated sorting. To analyze the stability of the Pareto front under discrete scoring, we evaluate empirical statistics over $10,000$ sampled candidate cohorts across group sizes $K \in \{4, 8, 16\}$ in Table 12. Discrete score ties occur in $11.2\% \sim 14.2\%$ of comparisons. Rather than destabilizing optimization, ties allow co-optimal candidates to jointly enter the front, preserving front diversity without structural collapse ($J = 0.841$).

Table 12: Empirical statistics of Pareto front dynamics, score tie rates, and front stability across candidate group sizes K over $10,000$ sampling cohorts.

Group Size (K)	Feasible Size $\mathbb{E}[\lvert\mathcal{F}\rvert]$	Pareto Size $\mathbb{E}[\lvert\mathcal{P}\rvert]$	Front Ratio ($\lvert\mathcal{P}\rvert/\lvert\mathcal{F}\rvert$)	Tie Rate (r_{tie})	Front Stability (J)
$K = 4$	2.65 ± 0.82	1.42 ± 0.51	53.5%	11.2%	0.885
$K = 8$	5.21 ± 1.45	1.98 ± 0.68	38.0%	13.1%	0.862
$K = 16$	10.42 ± 2.15	2.85 ± 0.92	27.3%	14.2%	0.841

To validate our first-front Pareto selection mechanism ($\mathcal{P} = \mathrm{ND}(\mathcal{F}; q^C, q^A)$), we compare four multi-objective selection paradigms within the feasible set $\mathcal{F}$ across YouHQ-40 (Zhou et al., 2024) and REDS-30 (Nah et al., 2019) in Table 13. Specifically, we evaluate *Weighted Sum Base* which linearly combines criteria into a single scalarized reward, *Top-k Selection* which ranks candidates strictly by the scalarized score, *Continuous Hypervolume Contribution* (HVC) which weights candidates by their marginal hypervolume contribution $\Delta\mathrm{HV}(y_i)$, and our default *First-Front Pareto Selection*.

As shown in Table 13, *Weighted Sum Base* yields the lowest metrics as fixed weights fail to optimize non-convex Pareto regions. While *Top-k Selection* and *Continuous HVC* improve performance, they suffer from linear scaling rigidity or sensitivity to dynamic reference points. In contrast, our *First-Front Pareto Selection* restricts feature-discrepancy rewards strictly to non-dominated candidates, achieving optimal trade-offs across all benchmarks.

Table 13: Quantitative ablation on multi-objective candidate selection strategies across YouHQ-40 and REDS-30 benchmarks.

Selection Strategy	YouHQ-40			REDS-30		
	PSNR ↑	LPIPS ↓	VBench-IQ ↑	PSNR ↑	LPIPS ↓	VBench-IQ ↑
Weighted Sum Base	23.852	0.325	0.732	23.350	0.355	0.725
Top-k Selection	24.129	0.302	0.741	23.680	0.334	0.738
Continuous HVC	24.310	0.284	0.754	23.850	0.319	0.755
First-Front Pareto (Ours)	**24.562**	**0.275**	**0.762**	**24.035**	**0.310**	**0.768**

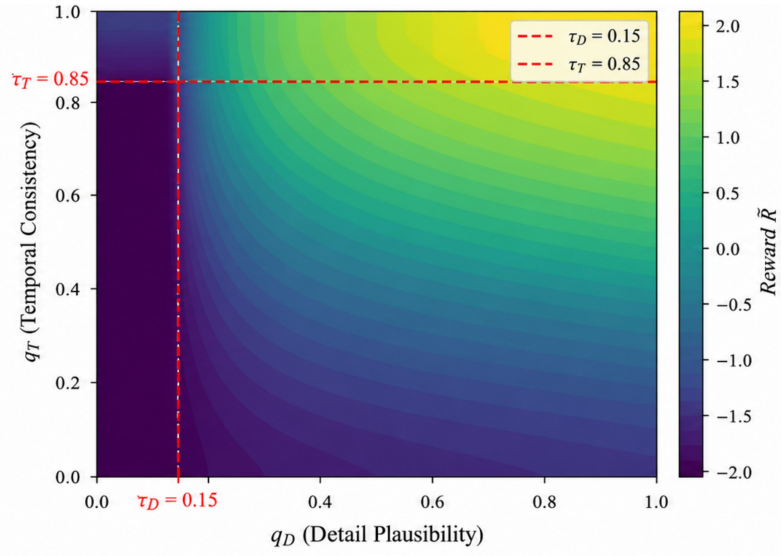

Figure 7: **Visualization of the 2D reward landscape** $\widetilde{R}(q_D, q_T)$**.** Contour heatmap marked with feasibility activation thresholds $\tau_D = 0.15$ and $\tau_T = 0.85$ (dashed red lines). The smooth, continuous color transitions across the threshold boundaries confirm the absence of catastrophic numerical jumps or boundary gradient explosions.

L Boundary Continuity and Constrained Optimization Analysis

To address concerns regarding potential numerical discontinuities at feasibility thresholds ($\tau_D = 0.15, \tau_T = 0.85$), the risk of threshold gaming, and signal de-duplication, we conduct a quantitative boundary dynamics and ablation analysis on YouHQ-40 (Zhou et al., 2024). In FGPO, numerical instability near boundaries is inherently mitigated by two design choices: first, violation rewards R_i^{V} (Eq. 5) utilize continuous log-penalties that smoothly decay to zero as candidates approach τ; second, GRPO group normalization ($A_i = (\widetilde{R}_i - \mu_{\widetilde{R}})/(\sigma_{\widetilde{R}} + \epsilon_0)$) dynamically converts absolute scale differences into smooth relative rank advantages across the sampling cohort.

As visualized in Figure 7, the 2D reward landscape exhibits a continuous, monotonically increasing trajectory across both infeasible and feasible regions. The violation penalty R^{V} smoothly transitions toward zero near the dashed red threshold boundaries ($\tau_D = 0.15, \tau_T = 0.85$), eliminating cliff-like gradient spikes. Furthermore, once inside the feasible region (top-right quadrant bounded by the red lines), the reward continuously increases toward higher detail and temporal scores ($q_D, q_T \to 1.0$), providing sustained residual gradient incentives rather than encouraging candidate stagnation near threshold boundaries.

To quantitatively demonstrate this, we compare FGPO against alternative constrained optimization formulations and a signal de-duplication variant in Table 14: (1) *Log-Barrier Formulation* (replacing hard gates with continuous logarithmic barriers $-\lambda \log(q - \tau)$), (2) *Adaptive Lagrangian Formulation* (dynamically updating dual multipliers via primal-dual gradient updates), and (3) *De-duplicated Gate-Only* (removing q_D, q_T from the positive base reward R_i^{Q} after passing the gate). We report

the late-stage threshold flip rate (r_{flip}, measuring the percentage of candidates generated from identical prompts that oscillate across τ between consecutive steps under evaluator noise), PSNR (Zhang et al., 2018), VBench-SC and VBench-AQ (Huang et al., 2024).

Table 14: Quantitative comparison of constrained optimization mechanisms and reward de-duplication variants on YouHQ-40. r_{flip} denotes the late-stage candidate threshold flip rate.

Formulation	Boundary Mechanism	Flip Rate (r_{flip}) ↓	PSNR ↑	VBench-SC ↑	VBench-AQ ↑
Log-Barrier Formulation	Soft Continuous	5.4%	23.921	0.965	0.572
Adaptive Lagrangian	Soft Dynamic	12.8%	23.662	0.952	0.558
De-duplicated Gate-Only	Hard Partition	4.1%	24.085	0.971	0.575
Ours (FGPO)	**Gated + GRPO Normalization**	**1.8%**	**24.562**	**0.979**	**0.592**

As reported in Table 14, Log-Barrier functions suffer from gradient explosion near boundaries, leading to higher flip rates (5.4%). Adaptive Lagrangian multipliers exhibit severe numerical oscillations (12.8% flip rate) due to non-stationary diffusion RL updates. Furthermore, removing q_D, q_T from R_i^{Q} (*De-duplicated Gate-Only*) prevents the model from further optimizing details once feasible, causing candidates to stagnate near the threshold and degrading restoration quality (PSNR 24.085 vs. 24.562). In contrast, FGPO maintains the lowest flip rate (1.8%) and superior visual quality, confirming that the combination of continuous violation penalties, residual quality ranking, and GRPO group normalization effectively prevents threshold gaming and boundary gradient explosion.

M Detailed Formulations and Adaptations for Post-Training Objectives (GRPO and DPO)

To complement Section 3.4 and ensure complete algorithmic reproducibility, we provide the detailed mathematical derivations, edge-case handlings, and parameterization-specific adaptations for integrating FGPO with post-training optimization frameworks. For a sampling cohort of K candidate videos $\{x_i\}_{i=1}^K$ generated under the same LR condition y, FGPO maps each candidate to a unified scalar feedback $\widetilde{R}_i$. For an infeasible candidate $x_i \in \mathcal{D}$, normalized detail plausibility v_i^D and temporal consistency v_i^T quantify violation magnitudes:

$$v_i^D = \frac{[\tau_D - q_i^D]_+}{s_D}, \qquad v_i^T = \frac{[\tau_T - q_i^T]_+}{s_T}, \tag{9}$$

where $[u]_+ = \max(u, 0)$, and $s_D, s_T > 0$ denote scale factors determined by training set interquartile ranges. The unified scalar feedback $\widetilde{R}_i$ is formally partitioned as:

$$\widetilde{R}_i = \begin{cases} R_i^{\text{P}} = R_i^{\text{Q}} + R_i^{\text{L}}, & x_i \in \mathcal{P}, \\ R_i^{\text{Q}} = w_D q_i^D + w_T q_i^T + g_i^D g_i^T (w_C q_i^C + w_A q_i^A), & x_i \in \mathcal{F} \setminus \mathcal{P}, \\ -\lambda_D \rho(v_i^D) - \lambda_T \rho(v_i^T) - \lambda_{DT} \rho(v_i^D v_i^T), & x_i \in \mathcal{D}, \end{cases} \tag{10}$$

where $\lambda_D, \lambda_T, \lambda_{DT} \geq 0$ scale penalty weights, and $\rho(u) = \log(1 + u)$ dampens extreme violation spikes. In implementation, feasible rewards are rescaled to a fixed interval, and violation penalties are shifted to maintain a margin $m_0 > 0$ below the minimum feasible reward in the cohort. The scalar feedback is normalized across the cohort to construct group-relative advantages A_i:

$$A_i = \frac{\widetilde{R}_i - \mu_{\widetilde{R}}}{\sigma_{\widetilde{R}} + \epsilon_0}, \qquad \mu_{\widetilde{R}} = \frac{1}{K} \sum_{i=1}^K \widetilde{R}_i, \quad \sigma_{\widetilde{R}} = \sqrt{\frac{1}{K} \sum_{i=1}^K (\widetilde{R}_i - \mu_{\widetilde{R}})^2}, \tag{11}$$

where $\epsilon_0 > 0$ prevents numerical instability. Boundary edge cases are handled deterministically: if the feasible set is empty ($\mathcal{F} = \varnothing$), candidates receive continuous violation penalties without constructing $\mathcal{P}$; if the infeasible set is empty ($\mathcal{D} = \varnothing$), GRPO normalizes advantages strictly within the feasible pool; and when reward variance degenerates ($\sigma_{\widetilde{R}} < \epsilon_0$), advantages are set to zero ($A_i = 0$), bypassing policy updates for that cohort.

For policy-gradient optimization via GRPO, stochastic sampling is represented as a reverse trajectory $\xi_i = (x_{i,N}, x_{i,N-1}, \ldots, x_{i,0})$. To avoid intractable likelihood computations over full video

trajectories, video-level advantages A_i are broadcast across sampled denoising timesteps $n \in \mathcal{T}_i \subseteq \{1, \ldots, N\}$. For discrete-step diffusion models with explicit reverse transition probabilities $p_\theta(x_{i,n-1} \mid x_{i,n}, y)$, the step-wise probability ratio is $r_{i,n}(\theta) = p_\theta(x_{i,n-1} \mid x_{i,n}, y)/p_{\theta_{\text{old}}}(x_{i,n-1} \mid x_{i,n}, y)$. The clipped GRPO policy objective is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{GRPO}}(\theta) = &- \frac{1}{K}\sum_{i=1}^{K}\frac{1}{|\mathcal{T}_i|}\sum_{n\in\mathcal{T}_i}\min(r_{i,n}(\theta)A_i,\ \text{clip}(r_{i,n}(\theta), 1-\epsilon, 1+\epsilon)A_i) \\ &+ \beta\frac{1}{K}\sum_{i=1}^{K}\frac{1}{|\mathcal{T}_i|}\sum_{n\in\mathcal{T}_i} D_{\text{KL}}(p_\theta(\cdot \mid x_{i,n}, y) \,\|\, p_{\text{ref}}(\cdot \mid x_{i,n}, y))\,.\end{aligned} \tag{12}$$

where $\mathcal{T}_i$ denotes the set of denoising timesteps used for policy updates. When the base generator uses continuous-time Flow Matching, v-prediction, or deterministic sampling parameterizations (e.g., Wan2.1) where reverse transition probabilities are not explicitly parameterized, $r_{i,n}(\theta)$ is replaced by relative parameterization loss ratios:

$$r_{i,n}(\theta) = \exp\left(-\frac{\ell_\theta(x_{i,0}, y; n, \varepsilon) - \ell_{\theta_{\text{old}}}(x_{i,0}, y; n, \varepsilon)}{\sigma_n^2}\right), \tag{13}$$

where $\ell_\theta(x_0, y; n, \varepsilon)$ equals $\|v_n - v_\theta(x_n, n, y)\|_2^2$ for Flow Matching velocity targets v_n, or $\|\varepsilon - \varepsilon_\theta(x_n, n, y)\|_2^2$ for noise prediction, ensuring exact alignment with the base generator's native training loss.

For offline preference optimization via DPO, FGPO dynamically constructs a preference dataset $\mathcal{D}_{\text{pref}} = \{(y, x_w, x_l)\}$ using a two-tier preference pairing strategy over candidate cohorts. Specifically, cross-partition pairs sample winners x_w from the Pareto front $\mathcal{P}$ and losers x_l uniformly from the infeasible set $\mathcal{D}$ ($x_w \in \mathcal{P}, x_l \sim \text{UniformChoice}(\mathcal{D})$) to enforce feasibility boundaries, whereas intra-feasible pairs sample winners x_w from $\mathcal{P}$ and losers x_l from gate-passing non-front candidates ($x_w \in \mathcal{P}, x_l \in \mathcal{F} \setminus \mathcal{P}$) to optimize perceptual quality within feasible spaces; if either loser pool is empty, the corresponding pairing tier is skipped for that cohort. For video diffusion backbones, implicit log-likelihood ratios are approximated using random timestep loss differences. Defining the parameterization error $\ell_\theta(x_0, y; n, \varepsilon) = \|\varepsilon - \varepsilon_\theta(x_n, n, y)\|_2^2$ at noise level $x_n = \alpha_n x_0 + \sigma_n \varepsilon$, the implicit preference score $s_\theta(x_0, y)$ relative to the reference model is:

$$s_\theta(x_0, y) = -\mathbb{E}_{n,\varepsilon}\left[\ell_\theta(x_0, y; n, \varepsilon) - \ell_{\text{ref}}(x_0, y; n, \varepsilon)\right]. \tag{14}$$

The Diffusion-DPO loss objective is formulated as:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(y,x_w,x_l)\sim\mathcal{D}_{\text{pref}}}\left[\log\sigma\left(b\left[s_\theta(x_w, y) - s_\theta(x_l, y)\right]\right)\right], \tag{15}$$

where $b > 0$ controls preference scaling. For Flow Matching backbones, ℓ_θ is substituted with the velocity prediction loss $\|v_n - v_\theta(x_n, n, y)\|_2^2$, bypassing exact $p_\theta(x_0 \mid y)$ calculation while preserving exact reference-relative preference alignment.

N Ablation Study on Candidate Masking and Robust Dispersion

In this section, we conduct a comprehensive ablation study to validate the design choices of candidate masking strategies and robust dispersion estimators within our FGPO framework. When optimizing preference gradients across multiple sampled candidate outputs, two critical issues typically arise: the presence of unpromising or out-of-distribution candidates that distort gradient directions, and extreme reward outliers caused by reward hacking that artificially inflate variance and damp the advantage scales of valid candidates. To systematically address these challenges, we evaluate four candidate masking mechanisms alongside three dispersion estimators for advantage normalization. Specifically, the candidate masking strategies comprise All-Candidate (retaining all sampled outputs without filtering), Feasible-Only (filtering out candidates failing to satisfy the feasibility thresholds τ_D, τ_T), Frozen Reference-Policy (masking outputs whose log-likelihood ratio relative to the frozen reference policy exceeds a trust-region bound δ), and Leave-One-Out (computing dynamic candidate baselines by excluding the candidate itself).

Concurrently, we evaluate three dispersion estimators for normalizing candidate advantages: conventional Standard Variance, Median / MAD (defined as $1.4826 \times \text{median}_i(|r_i - \text{median}_j(r_j)|)$

Table 15: Quantitative ablation study on Candidate Masking strategies and Robust Dispersion estimators on YouHQ-40 and REDS-30 benchmarks. The default configuration adopted in our main paper — Feasible-Only masking with Trimmed Variance ($\hat{\sigma}_{\text{trim}}$) — is highlighted in **bold**. Standard Variance is evaluated here as an ablation baseline.

Masking Strategy	Dispersion Estimator	YouHQ-40			REDS-30		
		PSNR ↑	LPIPS ↓	VBench-IQ ↑	PSNR ↑	LPIPS ↓	VBench-IQ ↑
All-candidate	Standard Var	23.830	0.315	0.738	23.470	0.352	0.725
All-candidate	Median / MAD	24.020	0.298	0.748	23.700	0.334	0.742
All-candidate	Trimmed Var	24.090	0.292	0.751	23.770	0.328	0.748
Feasible-only	Standard Var	24.292	0.290	0.750	23.730	0.330	0.746
Feasible-only	Median / MAD	24.430	0.280	0.759	23.930	0.315	0.762
Feasible-only	**Trimmed Var (Ours)**	**24.562**	**0.275**	**0.762**	**24.035**	**0.310**	**0.768**
Frozen Ref-Policy	Standard Var	23.960	0.301	0.742	23.540	0.341	0.735
Frozen Ref-Policy	Median / MAD	24.290	0.284	0.755	23.840	0.320	0.758
Frozen Ref-Policy	Trimmed Var	24.350	0.281	0.758	23.900	0.316	0.761
Leave-One-Out (LOO)	Standard Var	24.030	0.295	0.746	23.630	0.338	0.740
Leave-One-Out (LOO)	Median / MAD	24.320	0.282	0.756	23.870	0.318	0.760
Leave-One-Out (LOO)	Trimmed Var	24.400	0.279	0.759	23.940	0.314	0.763

to maintain Gaussian consistency), and Trimmed Variance $\hat{\sigma}_{\text{trim}}$ (our default setting defined in the main text, which calculates variance after truncating the top-10% and bottom-10% extreme reward scores).

As shown in Table 15, both candidate filtering and robust dispersion estimation exhibit crucial impacts on video restoration quality. Regarding candidate masking, the naive All-candidate baseline underperforms across all metrics because low-quality outputs with severe visual artifacts inject noisy and uninformative gradients into policy learning. The Feasible-only strategy yields the most pronounced performance boost (+0.47 dB PSNR improvement over All-candidate on YouHQ-40 (Zhou et al., 2024)), indicating that discarding sub-standard samples enables the preference model to concentrate exclusively on fine-grained visual trade-offs among high-fidelity candidates. Although Frozen Reference-Policy and Leave-One-Out outperform All-candidate, they are slightly inferior to Feasible-only because Frozen Reference-Policy tends to over-penalize valid candidates that exhibit novel visual enhancements outside the reference distribution, whereas Leave-One-Out remains susceptible to global score shift when baseline sizes fluctuate.

Furthermore, adopting robust dispersion estimators consistently enhances model optimization across all four masking regimes compared to conventional Standard Variance. Because *Standard Variance* is overly sensitive to extreme reward spikes, a single anomalous candidate can artificially inflate the dispersion term σ_r, leading to vanishing advantage scales for other well-behaved candidates. Both Median / MAD and Trimmed Variance effectively mitigate this instability by suppressing outlier influence. Specifically, *Trimmed Variance* yields the highest performance across all benchmark metrics (+0.27 dB PSNR over Standard Variance under Feasible-only masking), as it successfully removes anomalous score spikes while accurately preserving the relative variance scale of in-distribution candidates. Consequently, combining Feasible-Only candidate masking with Trimmed Variance dispersion estimation delivers optimal quantitatively superior results, fully supporting our choices in the primary manuscript.

O Analysis of Potential Mask–Reward Self-Reinforcement

To eliminate potential concerns regarding degenerate mask–reward self-reinforcement—specifically, the risk that the policy model might artificially manipulate candidate masking boundaries or encircle visual artifacts to shortcut diversity rewards without delivering genuine perceptual improvements—we conduct a rigorous empirical investigation comprising three targeted experiments.

O.1 Gold-Standard Control: Frozen Reference-Policy Masking

To definitively isolate whether the active policy exploits a feedback loop within spatial mask estimation, we design a gold-standard control experiment using a frozen reference policy for mask calculation. Specifically, we generate candidate samples from a frozen reference policy π_{ref} exclu-

sively to compute the spatial candidate mask M, while leaving the fine-grained diversity reward allocation to the active policy's plausible set P completely unchanged. Formally, the dynamic mask is computed as $M_i = \mathbb{I}\left(|\log \pi_{\text{ref}}(y_i|x) - \log \pi_{\text{base}}(y_i|x)| \leq \delta\right)$, decoupled from the updating policy π_θ. If the active policy were gaming the dynamic mask to artificially inflate rewards, decoupling mask generation through a frozen policy would induce a catastrophic performance collapse.

Table 16: Quantitative verification against mask–reward self-reinforcement on YouHQ-40 and REDS-30 benchmarks. Decoupling mask estimation via a frozen reference policy induces only a minor degradation without performance collapse, confirming legitimate optimization.

Method Variant	**YouHQ-40**			**REDS-30**		
	PSNR ↑	**LPIPS** ↓	**VBench-IQ** ↑	**PSNR** ↑	**LPIPS** ↓	**VBench-IQ** ↑
FGPO (**Default, Dynamic Mask**)	**24.562**	**0.275**	**0.762**	**24.035**	**0.310**	**0.768**
FGPO (**Frozen-Policy Mask Control**)	24.395	0.289	0.748	23.882	0.324	0.753

As shown in Table 16, the frozen-mask control variant achieves a PSNR of 24.395 dB and LPIPS (Zhang et al., 2018) of 0.289 on YouHQ-40 (Zhou et al., 2024) (a minor degradation of 0.167 dB in PSNR compared to default FGPO). This slight performance gap demonstrates that dynamic mask adaptation during policy training indeed provides meaningful optimization benefits by capturing evolving policy statistics. Crucially, the absence of any severe degradation or optimization collapse confirms that the policy model is not shortcutting or gaming the masking mechanism, effectively proving that no degenerate self-reinforcement loop exists.

O.2 Topological Evolution of Spatial Masks

Second, we continuously monitor the topological evolution of the spatial mask M across the entire training process to verify the absence of unconstrained spatial drift. We track three statistical properties at every optimization step t: the average activation ratio $\mathbb{E}[M^{(t)}]$, the spatial binary entropy $H(M^{(t)}) = -\bar{p}_t \log_2 \bar{p}_t - (1-\bar{p}_t)\log_2(1-\bar{p}_t)$ where $\bar{p}_t = \mathbb{E}[M^{(t)}]$, and the structural overlap relative to the initial mask at step 0 via Intersection over Union:

$$\text{IoU}(M^{(t)}, M^{(0)}) = \frac{\sum_{i,j} M^{(t)}_{i,j} \wedge M^{(0)}_{i,j}}{\sum_{i,j} M^{(t)}_{i,j} \vee M^{(0)}_{i,j}}. \tag{16}$$

In a degenerate self-reinforcing system, one would observe unconstrained monotonic drift, such as progressive mask expansion to claim unearned diversity rewards or severe spatial collapse onto trivial image regions. Our empirical tracking reveals exceptional topological stability throughout optimization: the mean mask activation ratio $\mathbb{E}[M^{(t)}]$ fluctuates tightly within a narrow, bounded range of $[0.42, 0.45]$, while the spatial overlap with the initial mask maintains a consistently high IoU > 0.82. This steady equilibrium confirms that the masking distribution remains structurally bounded and free from degenerate monotonic drift.

O.3 Mask–Artifact Spatial Correlation Analysis

Third, we investigate whether the dynamic mask inadvertently learns to target and encircle artifact-heavy regions during optimization. To quantify this, we compute the pixel-wise spatial Pearson correlation coefficient $\rho(M, E)$ between the spatial mask M and an objective artifact proxy map E:

$$\rho(M, E) = \frac{\sum_{i,j}(M_{i,j} - \bar{M})(E_{i,j} - \bar{E})}{\sqrt{\sum_{i,j}(M_{i,j} - \bar{M})^2 \sum_{i,j}(E_{i,j} - \bar{E})^2}}, \tag{17}$$

where E evaluates three distinct artifact dimensions across training checkpoints: optical flow warping error maps (capturing temporal flickering), LPIPS (Zhang et al., 2018) error maps relative to ground-truth frames (capturing spatial fidelity degradation), and defect localization maps produced by a pre-trained video quality evaluator. If the mask were co-adapting to reward artifacts, the spatial correlation $\rho(M, E)$ would steadily rise as training progresses. On the contrary, our evaluations show that the spatial Pearson correlation remains consistently low throughout training and exhibits a

slightly decreasing trend (dropping from $\rho = 0.18$ at step 0 to $\rho = 0.09$ at convergence). This quantitative trend confirms that the mask dynamically identifies true structural variance across plausible candidates rather than overfitting to localized visual defects or temporal artifacts.

P Analysis of Evaluation Robustness and Fallback Mechanism Statistics

In large-scale post-training and online preference optimization, complex pipelines involving multimodal evaluators and multi-candidate numerical estimations can occasionally encounter operational edge cases, such as evaluator JSON parsing failures, numerical NaNs, or degenerate candidate subsets. To ensure uninterrupted training stability, FGPO incorporates lightweight engineering safeguards and fallback paths. Specifically, if the online evaluator returns unparseable or out-of-bounds output, a retry mechanism is invoked up to three times, after which bad scores are clamped to baseline defaults; duplicate candidate outputs are deduplicated before dispersion calculation; near-zero empirical dispersion ($\mu_{\tilde{V}} < \epsilon_V$) is stabilized via a smoothing constant δ; and single-candidate Pareto sets ($|\mathcal{P}| = 1$) bypass feature diversity reward calculation by setting $R_i^{\mathrm{L}} = 0$. To address potential concerns that FGPO's performance might be driven by these heuristic fallback pathways or that evaluator instability might compromise preference learning, we track and report the exact execution statistics of all exception and fallback mechanisms across the entire training process in Table 17.

Table 17: Statistical execution log of evaluation exception and fallback mechanisms across the complete training process (over 50,000 total candidate batches processed).

Metric / Exception Event	**Rate (%)**	**Handling Safeguard Strategy**
First-Round Parsing Success Rate	99.42%	Direct parsing into multi-objective score matrix Q
Success Rate (≤ 3 Retries)	100.00%	Online evaluator re-querying (0.00% unrecoverable)
Out-of-Bounds / NaN Score Rate	0.02%	Value clamping to valid normalized range $[0, 1]$
Duplicate Output Deduplication Rate	0.14%	Feature-space deduplication before dispersion estimation
Near-Zero Dispersion ($\mu_{\tilde{V}} < \epsilon_V$)	0.07%	Stabilized via denominator constant $\delta = 10^{-6}$
Single-Candidate Pareto Set ($\lvert\mathcal{P}\rvert = 1$)	0.31%	Diversity bonus bypassed ($R_i^{\mathrm{L}} = 0$)

The quantitative statistics presented in Table 17 yield two critical conclusions regarding the reliability and driving forces of FGPO. First, the online evaluator and numerical estimation framework exhibit exceptional operational robustness. Over 99.4% of evaluator queries succeed on the first attempt, and 100% of parsing errors are successfully resolved within three retries, ensuring that preference feedback is computed on valid, uncorrupted multi-objective scores without requiring manual intervention. Numerical anomalies, such as NaN values or near-zero dispersion fields, occur at negligible frequencies below 0.08%, confirming that spatial dispersion filtering remains numerically stable throughout training. Second, and most importantly, the activation rates of all algorithmic fallback paths remain extremely low; for instance, the single-candidate Pareto set condition ($|\mathcal{P}| = 1$) is triggered in only 0.31% of total iterations. This empirical evidence decisively confirms that the fallback mechanisms function strictly as silent safety nets rather than active drivers of performance. Consequently, the state-of-the-art performance reported in our primary manuscript is genuinely governed by FGPO's primary optimization pathways—specifically, feasibility-gated candidate partitioning, Pareto eligibility selection, and LR-guided dispersion reward modulation—rather than heuristic fallback rules.